End of 2021 Progress Report - Multiomics Provider

1. Overview - Program-level milestones

During the Translator Phase II (Development), our Multiomics Provider team analyzed and curated real-world evidence from multiomics data and from electronic health records, generated knowledge graphs (KGs) representing significant connections observed in the data, exposed these KGs to the Translator ecosystem in multiple ways, participated in and contributed to multiple workgroups, and directly collaborated with other teams to support the design, implementation and testing of the Translator system.

We contributed to program-level milestones in multiple ways:

- Registered our KPs in SmartAPI.
- Collaborated with the SmartAPI team to register and continually update our KPs.
- Wrote generalizable parsers that transform each Multiomics KGX (EHR, Wellness, and BigGIM II) from our data-analysis knowledge graphs into a biolink compatible format that can be processed by Service Provider and exposed as a SmartAPI. We have collaborated closely with both Service Provider and Biolink to ensure that our components meet Translator Release Process requirements well in advance of deadlines, and helped coordinate timing of integration of features into our KGs and changes in Translator-wide standards.
- Provided detailed documentation on all KPs in the NCATSTranslator GitHub repository.
- Developed a testing plan to verify validity of knowledge graphs generated from the real-world data, service availability and knowledge retrieval. Implemented domain-agnostic and domain-informed QC of KGs and selection of representative KG subsets for efficient testing. Conducted extensive Translator-wide testing through design, implementation of use cases and Demos, identifying and helping resolution of over 100 issues outside of our KPs. Collaborating with Service Provider and BTE for future automated end-to-end testing in test and production environments.
- Developed and implemented a User Engagement Plan. We met with eight different scientists (seven outside of Translator) from five different institutions and areas of research, and conducted nine user engagement sessions. In each case, we started with broad open ended questions, listening to what challenges each SME was most interested in having Translator help solve. This highlighted some current limitations of Translator that could be addressed in the future, and provided several fruitful use cases to help drive testing and demos. Early on, the work primarily helped reveal new bugs and performance issues across Translator, which we drove to resolution. Over time, as Translator has become more performant, functional and stable, we were able to dive into increasingly complex real-world queries. This work led directly to creating new demos that provided genuinely interesting results for SMEs, and a series of use cases to help prioritize goals for 2022.
- Helped drive Translator-wide success for the December Demo:
    - *Workflow A:* From the use cases raised in the precision medicine section from May 2021 Relay, we reported how our BigGIM II KP are contributing to answer these questions, especially for the RHOBTB2 use case. We reported our findings using the BigGIM II Drug response KP. As RHOBTB2 is a candidate tumor suppressor gene, which is not easily to be targeted, there are currently no FDA approved drugs that can target RHOBTB2. We then infer from other diseases (cancers) about how the alteration of the gene is associated with sensitivity to certain drugs or chemicals using **Multiomics BigGIM II - Drug response KP**, and we found that JAK inhibitor Ruxolitinib, MAP2K5 inhibitor XMD8-85 shows higher sensitivity to certain cancer types. Using **Multiomics**

BigGIM - Gene Gene Interaction KP, we inferred from the gene gene interactions or regulations, and got the list of genes which show significant correlation with RHOBTB2, and targeted by FDA approved drugs. It provides new clues for users to explore new possibilities for testing and evaluation. This case was further derived as workflow A use case for December Demo.

- *Workflow B:* Provided medical knowledge to improve breadth of hepatic (liver) conditions included in queries. **Multiomics EHR KP** contributes query results.
- *Workflow C:* We combined two approaches to developing a compelling demo: 1) SME User Engagement to hear what real world problems people wanted to solve and 2) proactively collaborating with KP and ARA teams to understand the strengths and limitations of their Translator features. We identified and drove resolution on a large number of issues across Translator and developed an initial working Demo showcasing multiple KPs and ARAs. We subsequently worked closely with SMEs to improve the value of queries and the demo narrative, navigated carefully around Translator limitations and frequently-changing responses. The final results highlighted a wide range of KPs and ARAs and provided valuable results for SMEs.
- *Workflow D:* Provided alternate knowledge graphs format for Improver ARA consumption, supporting multiple questions in scenario D with **Multiomics Wellness KP** and **Multiomics EHR KP**.

2. The Multiomics KPs

The KGs we developed (see **Table 1**) contribute important, unique value to current use cases and lay the foundation for improvements and new KGs we will produce in subsequent years.

Table 1. *Knowledge Graphs produced by Multiomics Provider*

Name	**Description**	**Node types and *CURIEs***	**Edge types(Biolink: Predicates)**
Big GIM II			
BigGIM II - Gene Gene interaction (expr-expr)	KG from public multiomics datasets in healthy tissue types (GTEx) and cancers (TCGA) to describe the co-expression of two genes in different contexts.	Gene expression (RNA product)	Gene to gene coexpression association (Biolink:correlated_with->negatively correlated with) (Biolink:correlated_with->positively correlated with) biolink:coexpressed_with
BigGIM II - Gene Gene interaction (mut-expr)	KG from TCGA dataset to answer which gene expression are associated with gene mutations in different cancers	Gene mutation Gene expression *MONDO, SYMBOL*	Gene mutation to gene expression associations (Biolink:associated_with_expression_of)
Big GIM II: Drug Response KG (gene mutation	KG from the GDSC dataset to answer which drugs show higher	Gene (mutation), Drug (Chemical)	Gene (mutation) to chemical association

based drug response)	sensitivity or resistance to samples with the mutation of one gene	*CHEBI, CHEMBL, ENSEMBL, NCBIGene, PUBCHEM, SYMBOL*	Gene (mutation) to drug association (associated_with - > associated_with_sensitivity_of associated_with_resistance_of)
BigGIM II - Drug response KG - gene expression based drug response	KG from the GDSC dataset to answer which drugs show higher sensitivity or resistance to samples with the expression of one gene	Gene (expression), Drug (Chemical)	Gene (expression) to chemical association Gene (expression) to drug association (correlated_with->negatively correlated with) (correlated_with->positively correlated with) (associated_with - > associated_with_sensitivity_of associated_with_resistance_of)
BigGIM II - Drug - Target	KG from the public available knowledge resource such as Drugbank	Gene, Drug	Drug to gene association
Big GIM II: TCGA Mutation Frequency KP	KG from TCGA dataset to describe which genes are associated with each tumor type	Disease, Gene	Associated with biolink:GeneToDiseaseAssociation->biolink:gene has variant that contributes to disease association
Wellness Multiomics			
Wellness Multiomics KP	KG from ISB Wellness data on clinical labs, proteins, metabolites	ClinicalFinding, Protein, Metabolite, MolecularActivity *CAS, CHEBI, HMDB, KEGG.COMPOUND, KEGG.DRUG, LOINC, MESH, PUBCHEM, UniProtKB,*	correlated_with, related_to

		KEGG.ORTHOLOGY	
EHR Clinical Risk KP			
EHR KP	KG from 11,000,000 EHR records for risk factors for predicting future disease/drug and for classifying likelihood of current disease/drug	SmallMolecule, DiseaseOrPhenotypicFeature *CHEBI, HP, MONDO, NCIT, RXCUI, SNOMEDCT, UNII*	associated_with_risk_for negatively_associated_with_risk_for
COVID Multiomics KP			
COVID Multiomics KP (discontinued)	KG from ISB's INCOV study	Protein	Protein to protein coexpression association
	KG from ISB's INCOV study	Protein, Disease	Protein to disease severity association (beyond the current biolink model)
BIG GIM I			
Big GIM I (legacy)	Remains available as service for other KPs		

Note, our endpoints are here:
https://github.com/NCATSTranslator/Translator-All/wiki/Multiomics-Provider

2.1 Multiomics Big GIM II (Milestones M1.7, M2.1, M2.6)

Knowledge graph development:

In Big GIM II, we have extended the concept of BigGIM to understand the regulation network from genomes which covers the levels of entities including disease, gene mutations, gene expression, and drugs. We applied statistical models or machine learning approaches to extract knowledge graphs (KG) from the large cohort of public data resources. The BigGIM II includes a set of knowledge graphs which cover the associations, interactions among genes in multi-omics level as well as their associations with drugs and disease.

2.1.1 BigGIM II - Gene Gene interactions (expression based)

BigGIM II - Gene Gene interactions (expression based) is an updated version for BigGIM I with updated datasets from tumor based co-expression or tissue based gene expression.

Update from the tumor based co-expression: With the updated datasets from TCGA pancan study, we used the new version of gene expression value from the ISB-CGC PanCancer Atlas BigQuery Tables (pancancer-atlas.Filtered.EBpp_AdjustPANCAN_IlluminaHiSeq_RNASeqV2_genExp_filtered) to generate the graph (BigGIM II - Gene Gene interaction (expr-expr)). Gene co-expression correlations were computed using Pearson correlation. Gene expressions with observations in

at least 25 samples were taken into consideration. Coefficient and p-value were derived from Pearson correlation analysis. The gene co-expression graphs in different tumor types are currently available.
Update from the tissue based co-expression: Previous BigGIM I version uses the gene expression datasets from GTEx (V6). Here, we updated the source data to GTEx (V8).Gene expression data (Gene TPMs) for different tumor types were downloaded from https://www.gtexportal.org/home/datasets (GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_tpm.gct.gz). Pearson correlation is used to analyze the co-expressions. Tissue types that include Blood, Brain, Adipose Tissue, Muscle, Blood Vessel, Heart, Ovary, Uterus, Vagina, Breast, Skin, Salivary Gland, Adrenal Gland, Thyroid, Lung, Spleen, Pancreas, Esophagus, Stomach, Small Intestine, Prostate, Testis, Nerve, Pituitary, Liver, Kidney, Cervix Uteri, Fallopian Tube, Bladder, and Bone Marrow were included. A total number of 17382 samples were used. We have developed a python package to accelerate the analysis process (https://test.pypi.org/project/BDEx/), which makes the knowledge graph generating process much faster and reproducible. Current graphs can be found in the google project bucket (multiomics_provider_kp_data/BigGIM/GTEx_co_expr). As these types of graphs provide a large amount of results, a more integrated approach will be needed to make full use of them more efficiently.

2.1.2 BigGIM II - Gene Gene interactions (mut-expr)

We have analyzed the association between gene expression and gene mutation using the multiomics data from cancer to understand the regulation between gene mutation and gene expression by statistical modeling from the mutation data and gene expression data for each tumor type from patient derived samples. We have generated a KG about gene mutation ↔ gene expression associations. We have made these results available through BigQuery Tables: *translatordevkps:Mut_dep_Expression_pancancer*.
We also used the cancer cell lines to extracted the gene gene interactions, eg. gene mutation ↔ gene expression associations using the newly developed tool BDEx. The KGs are available at https://storage.googleapis.com/multiomics_provider_kp_data/BigGIM/Mut_dep/CCLE_mut_dep_expr_merged.Dec.2021.csv. We will further apply the TRAPI standard to format, and make updates of the BigGIM II KGs by adding the new results. We will implement them into full KP by updating the current BigGIM II APIs.

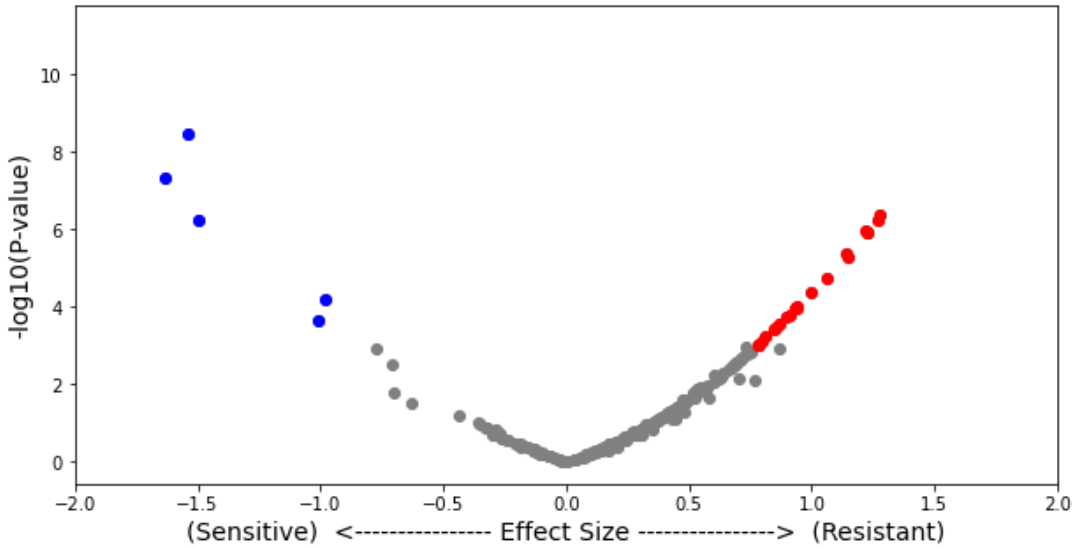
2.1.3 BigGIM II - Gene Gene interactions (Signaling interaction)

We also introduced gene regulatory networks by extracting graphs from other knowledge resources and literatures. The integration of the general gene regulatory work and the data driven tissue specific or disease specific gene gene interactions will provide disease specific gene gene interaction, which can be used for disease specific drug prediction or target prediction etc. We have also parsed the gene regulatory interaction from the KEGG pathways to extract the gene gene interaction. The edges include concepts of “activation”, “inhibition”, “binding” etc. The new KG can be downloaded from https://storage.googleapis.com/multiomics_provider_kp_data/BigGIM/Signaling/Signaling.csv. A biolink-compatible graph is still in progress.

2.1.3 BigGIM II - Drug response KG (mutation based)

To understand how different gene mutations are associated with different drug responses, we generated Multiomics DrugResponse KG (mutation based). Whole Exon Sequencing data and drug screening data from GDSC study (Iorio et al., 2016, Cell 166, 740–754) were used for knowledge graph extraction. Based on the mutation status of each gene, we grouped the cell lines into wild type groups and mutated groups. The significance of the difference of the drug

response IC50 values between the two groups was tested using Student T-test. The Drug response KG connects gene mutation and drugs in different disease contexts (26 tumor types). Each connection is annotated with evidence (Effect size), confidence (P-value), context (disease type), etc. The KG can provide evidence to address the question of which drugs are associated with a disease in a specific genetic background.
We established the Multiomics DrugResponse KP by collaborating with the Service Provider team; the KP can be accessed through https://biothings.ncats.io/drug_response_kp. The example query jupyter notebook (**Figure 1**) can be found at https://github.com/gloriachin/BigGIMII_API/blob/master/Notebooks/Query_MPKGs_DrugResponse.ipynb. We have further extended the Drug response KG by incorporating new datasets, such as BeatAML[PMID: 30333627], which measures the genetic alteration and and drug screening for each patient derived sample. We make an updated version of the Drug response KG by incorporating the added KG.




```
result_sig_sensitive.sort_values(by = ['edge_confidence_p'])
```

subject_id	subject_symbol	subject_type	object_id	object_name	object_type	edge_label	edge_context_disease
1956	EGFR	Gene	PUBCHEM:57519523	Afatinib	ChemicalSubstance	related_to	MONDO:0005061
1956	EGFR	Gene	PUBCHEM:57519523	Afatinib	ChemicalSubstance	related_to	MONDO:0005061
1956	EGFR	Gene	PUBCHEM:5328940	Bosutinib	ChemicalSubstance	related_to	MONDO:0005061
1956	EGFR	Gene	PUBCHEM:123631	Gefitinib	ChemicalSubstance	related_to	MONDO:0005061
1956	EGFR	Gene	CHEMBL.COMPOUND:CHEMBL1201577	Cetuximab	ChemicalSubstance	related_to	MONDO:0005061
1956	EGFR	Gene	PUBCHEM:65110	AICAR	ChemicalSubstance	related_to	MONDO:0005061

Figure 1. *Example query results from the DrugResponse KP. This example shows which drugs show higher resistance (red) or sensitivity (blue) in patients with lung adenocarcinoma (MONDO: 0005061). The top drugs are shown in the lower panel.*

2.1.4 BigGIM II - Drug response KG (expression based)

To understand how different gene expressions are associated with different drug responses, we developed BigGIM II- Drug response KG (expression based). Spearman correlations were calculated between the gene expression (RMA gene expression values) and drug response Area Under the Curve (AUC) for cell lines in different tumor types in the GDSC project.
The correlations were calculated only if the number of cell lines with both the drug response data and gene expression value for more than 6 samples. For each tumor type, the correlations between **gene** (symbol), **drug name** (need to transform into drug ids), **correlation**, **p-value**,

sample size and **tumor types** are included. The Drug response KG (expression based) can be used to answer the expression of which gene is associated with sensitivity or resistance to which drugs or chemicals.
The drug response KG represents multiple providence, evidence, as well as different qualifiers to the genes or drug response. By working with the SRI team and the Biolink team, we define an example of new schemas with modifiers and qualifiers, which could provide more clearer information for further integration. To make the knowledge graph generalizable, by working with the SmartAPI team, we connect our KP with the SmartAPI team by writing a standardized parser to connect our KGs to their standardized APIs (see Deployment section).

2.1.5 BigGIM II - Drug - Target

We have parsed the DrugBank database [1], and extracted the drug ↔ gene associations, to provide the current knowledge about the gene or gene product as the target for FDA-approved anticancer drugs. We have made this KG available through BigQuery Table: translatordevkps:DrugBank_Drug_Target. We further extract the drug-target interaction from DrugCentral [2], Therapeutic target database [3], and other literature [4]. It currently includes 71,280 interactions between drugs and targets from multiple resources. The KG is currently available in the project google bucket: https://storage.googleapis.com/multiomics_provider_kp_data/drug_response/CTDMB_formated.csv, further updating of the BigGIM II by adding this graph is still in progress.

2.1.6 Big GIM II - TCGA Mutation Frequency KG

We extracted gene mutation frequency from different tumor types from the pan-cancer atlas on the ISB-CGC platform, and produced KG about TCGA_mutation_frequency KG with nodes of "Disease" and "Gene mutation", and edges of "associated with" associations to understand which gene is considered as a driver or highly frequently mutated in specific contexts (disease or tumor type). The evidence for each connection is weighted by Frequency. This KG covers 33 tumor types and all possible gene mutations, so users can filter the frequency at their own preference. An API was also established to query the KG, which can be accessible from https://biothings.ncats.io/tcga_mut_freq_kp. Example query jupyter notebook: github.com/gloriachin/BigGIMII_API/blob/master/Notebooks/Query_MPKGs_Mut_Freq.ipynb. We have demonstrated the Drug-dependency KG and TCGA_mut_freq KG at the September Relay meeting. By working together with the Service Provider, Exploring Agent, Explanatory agent, Clinical Data Services Provider and the (im)Prove team, we were able to answer the questions on drug selection for different tumor types according to the mutation status or the disease types.

Tool development:
Tools for knowledge graph extraction: As most of our knowledge graphs will need intensive analysis from large data resources, to make the results reproducible, we have developed tools for the extraction of publicly available datasets, performing statistical testing in high efficiency by developing optimized functions. We are developing a python library (BDEx: Biomedical data Exchange: https://test.pypi.org/project/BDEx/), including functions for efficient computing of Spearman correlation and Ranksum test. It can reduce the computing time from days to hours for one set of gene expression matrices for a single laptop.
Tools for knowledge graph query: As part of the testing plan, we also developed a testing environment for the querying of the knowledge graphs before exposing to the production version or registering in smartAPI. It is currently deployed in http://35.233.133.157:5000/docs. The BDEx python library being developed can also query the knowledge graphs from this endpoint,

which could provide a convenient way for testing the results and make integration of the KPs being developed.

Scientific study using BigGIM II:
With the accumulation of the knowledge graphs for gene-gene interaction, gene-drug interaction, disease-gene interaction etc, we are able to answer some translational questions. As a use case, we can predict targets for patients for specific types of genetic alterations, and develop further tools on top of the KGs [Fig 2], which may be further useful for the user-interface design and implementation. Here, we show an example of inferencing targets that sensitize drug resistance using an example of KRAS mutated patients. KRAS mutations comprise a large subset of lung cancer, but no directed targeted therapies are available that effectively control the KRAS activation induced by KRAS. Patients with KRAS-driven tumors are commonly under standard cytotoxic chemotherapy, which are often transiently effective due to drug resistance [PMID: 30171261]. The exploration of potential targets that can sensitize drug resistance induced by KRAS activation is crucial for better patient treatment. Here, using the knowledge graphs we developed, we predicted mTOR as a potential target that could sensitize drug resistance caused by KRAS mutation. It was supported by both the mutation based drug response KP, expression based drug response KP, and BigGIM interaction KP. Hyperactivated mammalian target of rapamycin (mTOR) pathway is a characteristic hallmark of KRAS-mutant lung adenocarcinoma after chemotherapy treatment, and that KRAS-mutant lung cancer cells rely on persistent mTOR signaling to resist chemotherapeutic drugs [PMID: 30171261]. Our results highlight the potential of using mTOR as the potential targets for KRAS mutated patients, which is consistent with the literature reported results that mTOR as a mechanism of resistance to chemotherapy in KRAS-mutant lung cancer and validate a rational and readily translatable strategy that combines mTOR inhibitors with standard chemotherapy to treat KRAS-mutant adenocarcinoma, the most common and deadliest lung cancer subset [PMID: 30171261]. We also found other candidate targets that could sensitize the KRAS mutated associated drug resistance, such as ABL2, BCR, CHUK, JAK2, MAP3K2, MAP4, MAPK1, MAPK7, MAPK9, NEK7, RIOK1, RIPK1, TBK1, TAOK1 etc. Gene set enrichment analysis of the candidate targets shows these genes show overrepresentation in the pathways of MAPK signaling pathway, adipocytokine signaling pathway etc. Manuscript in preparation.

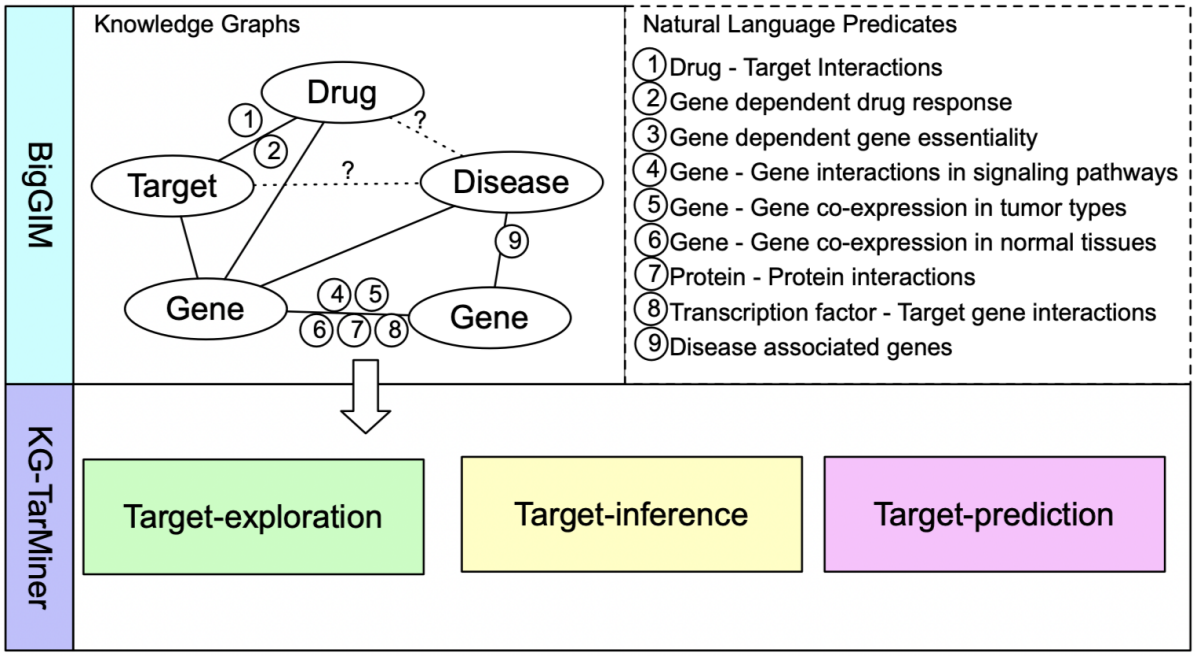


***Figure 2**. Overview of knowledge graph based Target Miner.*

2.2 Multiomics Wellness KP (Milestones M1.8, M2.2, M2.7)

Derivation of knowledge from wellness data

- We have analyzed the ISB Wellness dataset, which has been phenotyped extensively, affording many types of correlations and connections to be uncovered. Expanding on our original wellness study of 108 individuals [5], this deep phenotyping data set integrates many data types including WGS and/or SNP genotyping, clinical blood tests, salivary cortisol, weight and BMI, blood pressure, health assessments, provider notes, gut microbiome, blood metabolomics, blood proteomics, activity tracking, sleep tracking, and heart rate. The cohort includes 4,879 individuals with at least one blood draw via Arivale. Integrative analysis of this multidimensional data set is already leading to significant novel findings, e.g., on the connection between blood metabolites and the microbiome [6] and how this reflects aging [7]. The data were collected in longitudinal ‘snapshots’ that enable a more detailed analysis of data accrual and stability than a single ‘final’ data set view can.
- We have created and deployed multiple versions of the Multiomics Wellness KG. We computed correlations among attributes in the chemistries, metabolomics and proteomics tables in the ISB Wellness dataset. These attributes are clinical labs, metabolites, and proteins, respectively. Each attribute can either be a blood analyte or an index computed from one or more analytes. For example, the chemistries table has Albumin, Globulin and also the ratio of the two as three different attributes. The resulting KG includes statistically significant correlations from the inner join of the clinical labs, protein panels and metabolites, thus extending on the original version that included only correlations within each table.
- We performed a detailed curation of LOINC codes for the analytes that have significant correlations with other analytes for the attributes in the chemistries table, and modified the biolink concepts of a subset of nodes to ClinicalFinding to retain LOINC codes that best preserve the identity of a node.
- We transformed the resulting set of attributes and correlations into a KG (**Figure 3**) and expressed them in the KGX format. This format expresses KGs as two TSV files: one for the nodes with necessary columns for curies (a compressed URI that uniquely identifies a node)

and a Biolink model concept; and the other representing edges between these nodes, requiring the curies for the two end nodes, a Biolink concept for the relationship between the two nodes and similarly another relationship concept expressed in some standard ontology. In collaboration with Kevin Xin (Su Lab, Service Provider) we deployed this KG via BioThings API.

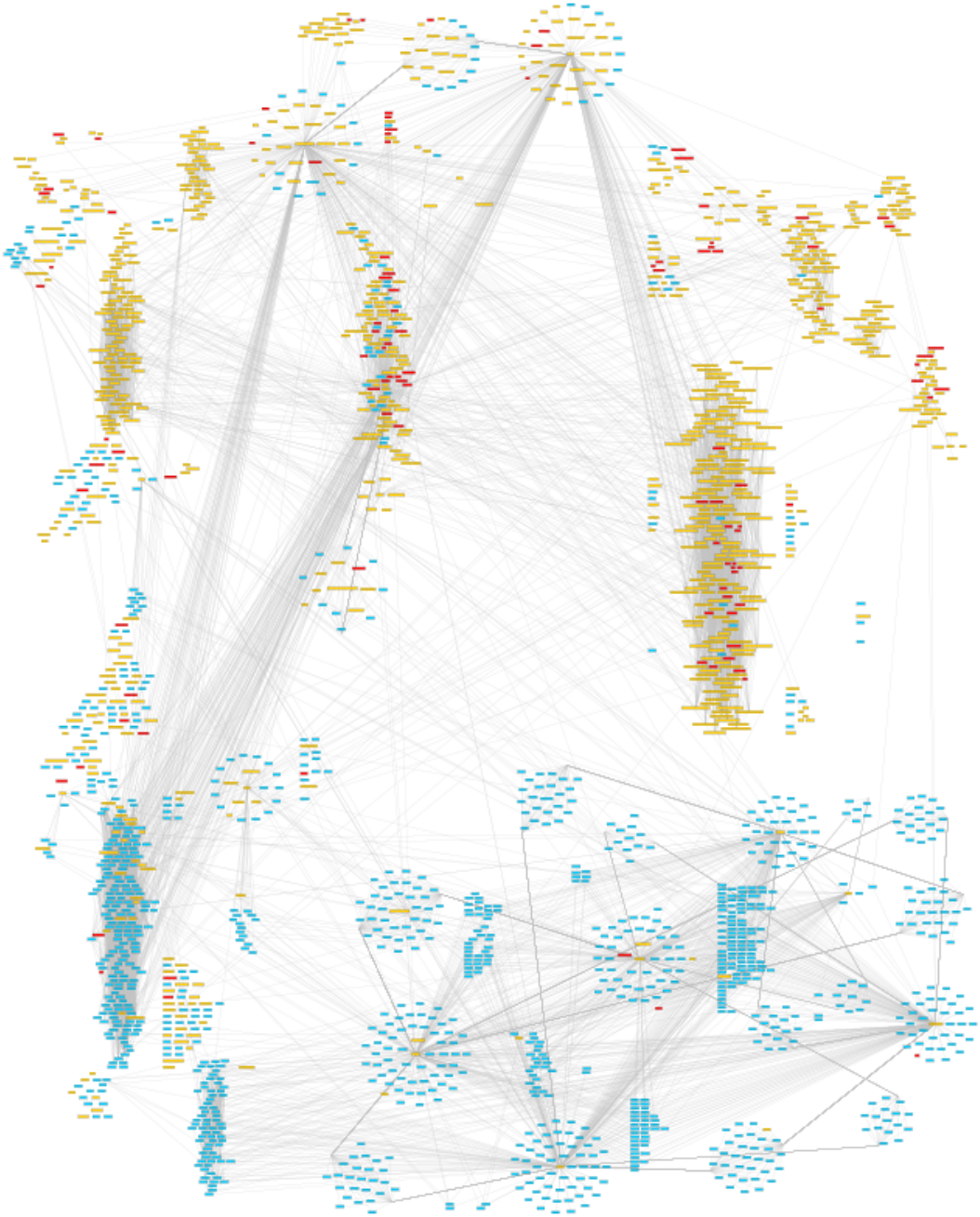


Figure 3. *Visualization of the core of the ISB Wellness Multiomics KG, v.1.3, integrating clinical chemistries (red), proteins (cyan) and metabolites (yellow), which includes drugs. The edges of the network are correlations between various analytes, retaining only edges with p-value under* 10^{-100}*. The complete network is much more extensive and complex than depicted, and subsequent versions (e.g., current v.1.6) even more so. Several metabolites are 'hubs' connected to multiple proteins; most of these hub metabolites are drugs.*

- Based on the analysis of snapshots (described below), we implemented metrics of knowledge confidence by taking into account the p-values of correlations between analytes as observed in former data snapshots. The information is presented in the form of a new column in the edges TSV called ‘weighted p-value’, the weight of the p-value in each snapshot is scaled by the factor of the number of observations in the corresponding snapshot. We will further refine this method by making significantly different snapshots (as assessed via data fingerprint comparisons) contribute more to the confidence in the derived knowledge.
- We have created the first version of a predictive model by creating a regression model for the analytes that are correlated with the most other analytes. As a result, we have a number of new edges and connections between analytes that are actually not significantly correlated with an analyte but are still a regressor of theirs. We used ridge regression for this purpose, so as to have a reliable prediction model for the number of baseline observations, which are fewer than the number of analytes. For example, the protein CVD3_O00300 (tumor necrosis factor receptor superfamily, member 11b) is not significantly correlated with adiponectin in the Wellness graph, but the regression analysis shows that this protein is one of the most important regressors (predictors) of adiponectin, consistent with the literature [8].
- The new edges so created have the predicate ‘related_to’ and a relation from the RO ontology that more closely describes the relationship of statistical prediction. Biolink is being updated to create a suitable predicate, possibly called ‘regressor_of’ as a hierarchical descendant of ‘predicts’ - see issue Biolink:731.
- We incorporated the results of stratification of correlations with additional context information on the individuals and the observations, like sex, age group, ancestry, seasonality, and proclivity to having extreme values for specific analytes.
- We updated the Wellness graph to adhere to the latest Biolink version to support results for the December demo.
- We created a new version of the Wellness graph with recomputed correlations between all analytes with a much higher concept pair count. We integrated results for the gut microbiome represented as molecular activity, in the form of its correlations with blood analytes and each other. There are ~5000 unique kegg orthologs that represent the gut microbiome analytes. We expect these results to be soon deployed by the Service Provider.
- We computed the *interactions* between various blood analytes and gut microbiome. For the analytes that were significantly correlated, we modeled them pairwise with other analytes to obtain a predictive relationship. We used *generalized linear model*, to obtain the interaction term (the coefficient of interaction) and its significance using the model: $analyte_1 \sim analyte_2 * analyte_3$ The interaction term $analyte_2 : analyte_3$ gives a measure of the change in relationship between $analyte_1$ and $analyte_2$ (or $analyte_3$) in the presence of $analyte_3$ (or $analyte_2$). Representation of such knowledge requires an association between three nodes - a limitation in the current Biolink Model.
- For example, among the interactions between metabolites and gut microbiome, one significant interaction found was between 1-methylnicotinamide, pyridoxate and glutamate carboxypeptidase [EC:3.4.17.11]. The interaction between pyridoxate and glutamate carboxypeptidase changes both in direction and magnitude of coefficient of relation between the three analytes:

$analyte_1$	1-methylnicotinamide
$analyte_2$	pyridoxate
$analyte_3$	glutamate carboxypeptidase [EC:3.4.17.11]
Coefficient $analyte_3$	-2.474
Coefficient $analyte_2$	-0.028
Coefficient interaction	1.78
P-value $analyte_3$	2.65e-05
P-value $analyte_2$	1.39e-16
P-value interaction	2.63e-91

- We encountered an extension of the 'interactions' paradigm when we explored the application of Differential Rank Conservation (DIRAC) [9] to ISB's wellness data. DIRAC requires two-fold separation within data. It provides quantitative measures of how network rankings differ either among networks for a selected phenotype or among phenotypes for a selected network. While the phenotypic separation can be extracted trivially from the Wellness dataset, for example sex or ethnicity based stratification, the network modules within data are hard to construe. The method is designed to analyze relative regulation of modules that are involved in biologically known processes among phenotypes.
- We applied DIRAC to the metabolomics data of ISB's wellness cohort. We divided the cohort into Males and Females, and identified five network modules to work with: 'Cofactors and Vitamins', 'Carbohydrates', 'Nucleotides', 'Energy' and 'Peptides'. The normalized rank indices of the metabolites that form these modules, among males and females are:

Network	**Females**	**Males**
Cofactors and Vitamins	243.8	200.9
Energy	33.6	34.4
Carbohydrate	97.9	85.8
Nucleotide	150.5	99.9
Peptide	544.3	411.6

These results indicate a biologically significant general trend of tighter regulation in the relative abundance of metabolites in most metabolic networks in females. As the Biolink model is adapted to support knowledge generated by graphs that contain N-ary relationships, we can expand the phenotypes for such analyses to contain rank conservation among network modules of diseases.

Analysis of snapshots in ISB wellness data to derive a knowledge stability metric

- Datasets that grow over time can have 'snapshots' (or 'freezes') stored periodically, for example to ensure that analyses done on the dataset are stable and reproducible, and don't

change with every small change to the dataset. The ISB Wellness dataset is such a dataset, with 96 snapshots spanning slightly over two years of data collection, and multiple data tables per snapshot. With each newer snapshot, the complexity of the dataset increased by adding timepoints for the same individuals, by adding new individuals, by adding (and sometimes, removing) attributes to existing tables, and even by adding new tables. The dataset could also change by removal of individuals, e.g., those that withdrew consent to participate in research.

- We sought to leverage the availability of the many dataset snapshots to derive metrics of stability or reliability of the knowledge derived from the data. Does confidence in knowledge derived from data increase as more data are added, or does it become weaker? Our preliminary results from correlations in the chemistries table suggests this can go both ways, and thus that there is value in performing this analysis.
- In some cases, consecutive snapshots may be very similar; some of the tables may be entirely unchanged. It would therefore be a mistake to take these to be independent when doing any statistical computation over the different snapshots, to assess the stability of edges in the knowledge graphs created from similar snapshots. This calls for an algorithm to assess the information content offered by a snapshot, and to prune [10] redundant snapshots, so as to increase the reliability of the stability metric.
- There is therefore a need for metrics of similarity between snapshots. Our data fingerprinting method [11] can efficiently yield such a metric. We computed data fingerprints for the ‘chemistries’ table in each of 96 snapshots of the ISB wellness data (gray-filled circles), and visualized using PCA (**Figure 4**). We observed a significant transition point followed by a faster rate of change at the beginning of 2019; a less pronounced shift in early 2018 corresponded to the addition of two columns to the table. Subsequent analyses explained the main (2019) effect as a result of changes in ‘reflexive’ measurements; excluding such measurements canceled the effect, and revealed a further minor shift in mid-2018 caused by the removal of two columns from the table.
- Insights from this analysis led to an improvement in the quality of the knowledge presented in the Wellness KG. Removal of the ‘reflexive’ measurements from the ‘chemistries’ table affected the derived knowledge. In one example, the statistically significant positive correlation between zinc_plasma_or_serum and triglycerides (which contradicts reports in the literature) was canceled by removal of the ‘reflexive’ measurements.
- We have a manuscript in preparation presenting these analyses, aiming to include similar analyses on other datasets with snapshot structure, to validate the generalizability of the method.
- We are currently working on adapting the code for the data fingerprinting algorithm [11] to be able to run it in the environment for EHR data. We have identified over 30,000 patients in the EHR data who have been diagnosed with IBD (Inflammatory Bowel Disease) and have lab results for them over a period of over 10 years. This information can be organized into snapshots similar to ISB’s wellness dataset. We sought to use the algorithms in [10] and [11] to derive similar metrics of stability for risk factors for IBD.
- We ported to Python the instance reduction algorithm [10] to make it platform independent. The new code is in Python which saves compute time and is easily refactorable.
- We evaluated additional datasets in snapshot format (e.g., PDB) for performing similar analyses.
- We expect this methodology to be applicable also to performing QC on versions of KGs stored in the Knowledge Graph Exchange (KGE) Archive.

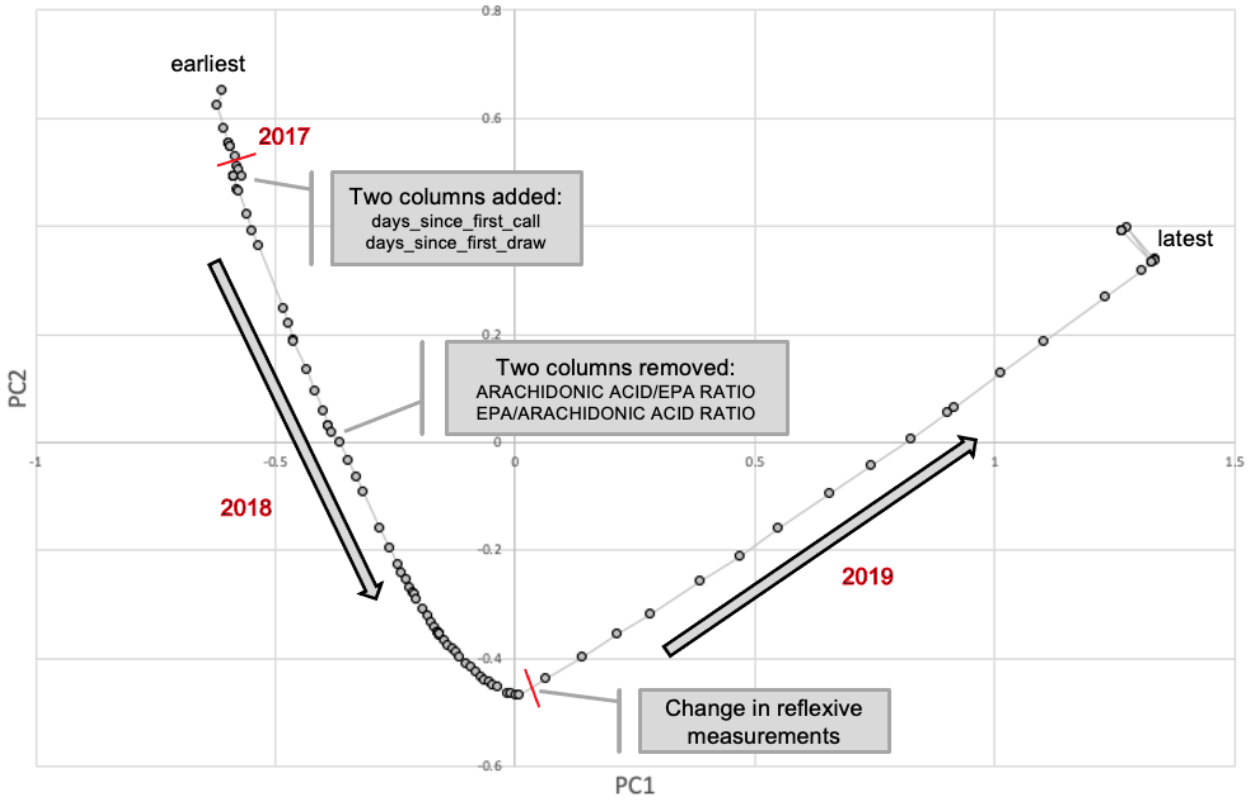


Figure 4. *Visualization of the trajectory of the ISB Wellness dataset (clinical chemistries table) over time.*

2.3 Multiomics EHR Clinical risk prediction KP (Milestones M1.9, M2.3, M2.8)

Derivation of potential risk factors and protective factors for phenotype, drug exposures and disease diagnoses, based on real-world evidence from EHR data for 20,000,000 patients.

- Initially, we developed the first biolink-compatible KP of predictive risk factors for the two major outcomes we had planned: long-term outpatient risk of sepsis and acute inpatient risk for developing serious COVID-19 illness. This work established foundations needed to scale to a wider breadth of phenotypes in the future: data extraction, clinical logic, integration of open source machine learning and analysis.
- Subsequently, we scaled up our KP to include 82 Disease or PhenotypicFeature nodes and 617 Drug or ChemicalSubstance nodes with over 10,000 edges. We stratified each patient's EHRs over the past 5 years into two intervals of 3 and 2 years. When building the models, we initially took features from the first 3 years and the outcomes from the next two years, thus preserving some temporal albeit noncausal information. We refer to this as "model A".
- To select the features for model A we first ran a chi-squared selector to filter out unimportant features. We trained on the logistic regression models, stratifying the data as described above, and chose the 100 or 50 most important features (depending on if the features were medications or diseases) to include as edges in the KP.
- Most recently, we have further scaled up the Multiomics EHR Risk KP to include 802 unique nodes (394 ChemicalSubstance, 196 Disease, 205 PhenotypicFeature, and 6 Procedure nodes) and 235,936 edges, using an improved modeling approach. **Figure 5** shows a portion of this knowledge graph. Each positive-feature-coefficient edge (indicating a positive or direct relation) has the (biolink:associated_with_risk_for) edge predicate. Each negative-feature-coefficient edge (indicating a negative or inverse relation) has the

(biolink:negatively_associated_with_risk_for) edge predicateWe currently specialize edge predicates by including parallel edges (biolink:correlated_with, biolink:negatively_correlated_with, biolink:related_to) and are engaging with Service Provider and SRI on alternate dynamic implementations.

- With this **>20-fold scaling of our Multiomics EHR Risk KP**, we were able to directly support a wide range of disease and drug queries featured in Just Fix It, Stand-up sessions and Demos.
- Additional contribution to multiple larger use cases listed in Section 5 below.
- To build our KP we ran multiple logistic regression models on a specified set of disease, medications, and labs. We binarize all of the variables, where 1 indicates the presence of a disease or medication and 0 indicates no such disease or medication. For the continuous laboratory results, we use the EHR-reported status to determine whether a particular result is high, low or neither. As such, we split each lab into two features: lab_high and lab_low where the specification of (1,0) or (0,1) indicates the lab result was high or low respectively, while "normal" (as defined by the reference ranges) and lack of lab result are mapped to (0,0).
- Each individual model can be represented by a node together with all the edges for which the chosen node is the child. The weights on those edges correspond to the feature coefficients of the logistic regression model. The predicates of these edges reflect the probabilistic interpretations afforded by the use of logistic regression models: the model coefficients are the rates at which the sigmoid functions change from 0 to 1. Negative coefficients indicate decreases risk, while positive indicates increased risk.
- As part of the Clinical Data Working group, we helped surface challenges and drive github issues with temporal logic, provenance, and context, which are essential for EHR real-world evidence and relevant for many other KPs. We have practical solutions for our KPs and are driving issues by contributing end-user use cases with ARAs. See use cases.
- Developed automated EHR-derived WHO severity outcomes for use for combined EHR + Multiomics modeling and scalable EHR risk prediction.
- We identified and addressed issues pertaining to interpretability of our edge weights: improving semantic precision of predicate names, and adding **log metrics** to provide more insight on prevalence/incidence, **while still preserving differential privacy**.

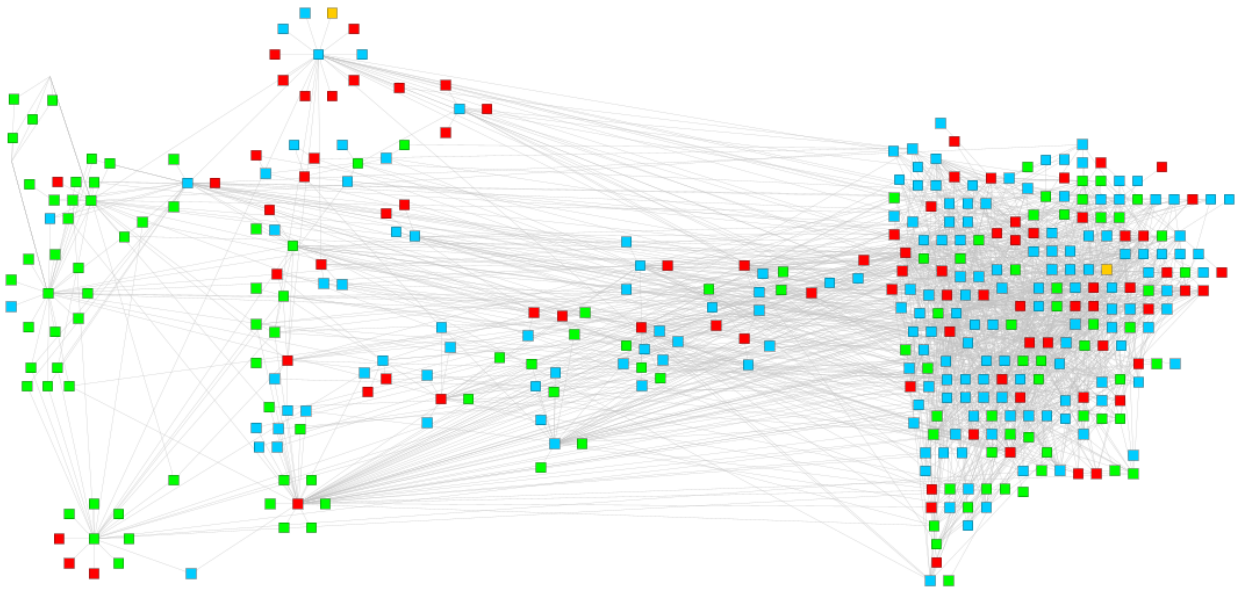


Figure 5. *Visualization of a subset of the EHR Multiomics KG, integrating chemical substances (red), diseases (cyan), phenotypic features (green) and procedures (yellow). The edges of the network are filtered for visualization, retaining only edges with absolute feature coefficient over 50. The complete network is much more extensive and complex than depicted.*

2.4 COVID Multiomics KP (Milestone M1.10, discontinued)

NOTE: This Milestone was discontinued at the end of 2020 per NCATS instructions.

- From the multi-omics data from 400 samples from COVID-19 patients in ISB's INCOV datasets, we investigated correlations of the molecular levels associated with serious illness and mortality in COVID-19. From the primary PCA analysis of the plasma protein expression levels [12], three groups of COVID-19 patients with different severity has been classified (mild group: WHO ordinal scale (WOS) = 1-2, moderate group: WOS=3-4, severity group: WOS = 5-7).
- Using the protein expression data measured using the olink platform, we analyzed the difference between the plasma protein expression levels among the three groups using the Kruskal-Wallis H-test, or the difference between any of the two groups using the Wilcoxon rank-sum test. Based on this data driven analysis, we generate the first version of the KG COVID19 severity KG, which features nodes of Proteins and Disease (COVID19), and edges of associated_severity_of disease.
- Using the COVID multiomics data, we further extended the Big GIM to other diseases, such as COVID19, to understand the co-expression of proteins in normal or COVID19 patients.
- Collaborated with other teams in the Translator community who are interested in COVID-19 to derive a more comprehensive knowledge for COVID-19. We drove and resolved biolink compatible approaches to support central clinical concepts around blood pressure. This included: observations, hypotension/hypertension relative to age, personal baseline, pregnancy status), hierarchical categories or hypotension/hypertension as diagnostic label, shock through proxies of interventions (drugs such as vasopressin, fluids), hypertension as a long term risk for serious outcomes, and relative hypotension as a short term risk.
- Papers published in Cell [12] and Clinical Infectious Diseases [13].

3. KP engineering (Milestones M1.11, M2.5, M2.9)

Testing

We have implemented our testing plan in three layers:

Layer 1: Preliminary KG evaluation. Since the KPs produced by our Multiomics Provider team involve knowledge derived from source data, a crucial first step is to evaluate the KG itself. We implemented a domain-agnostic QC method to evaluate each newly generated KG, expressed in KGX standard two-TSV format, for internal consistency. For each of the two TSV files, we perform multiple tests of data types included, repetitions of values, missingness, etc. We also evaluate the consistency between the two TSV files (e.g., whether declared nodes have no associated edges, and whether edges refer to undeclared nodes). All these tests are summarized in a compact JSON format report. We then evaluate any QC flags in the report and make a domain-informed assessment of whether they are justified (e.g., nodes can have no associated edges), or alternatively whether they reflect computational or representation failures (e.g., multiple nodes with identical identifiers, edges linking undeclared nodes). We also used well-known interactions for domain-informed verification. For example, for the BigGIM Drug Response KP, we use the drug-target interaction as a cross checking of the graph by examining whether the target gene itself is a predictor or biomarker for its targeted drugs. The comparison of evidence such as p-values etc between the new generated graph and well-known interaction will give us more confidence about the graph itself, as well as providing ways for graph ranking and results selection.

Layer 2: Internal querying. To perform testing of querying new KGs in-house, before deployment via BioThings Explorer (BTE), we implemented an internal queryable endpoint using fastAPI [http://35.233.133.157:5000/docs]. Since KGs can be very large, we implemented a procedure for selecting a representative set of significance-sorted edges that can be tested to assess the integrity of nearly the entire KG. For example, this procedure selected 939 edges out of the 229,614 edges in version 1.3 of the Multiomics-Wellness graph. This selection procedure is performed for each version of each KG. The testing queryable API can be used to test the time cost for a given query before introducing the KG into the SmartAPI. It also makes a comparable endpoint with the developed TRAPI endpoint for the integrity test.

Layer 3: External deployment. Once a KG is deemed suitable as per layer 1 and layer 2 testing, we deploy it via BTE, and perform testing in collaboration with the Service Provider team, and through querying via ARAs and via the ARS.

Deployment

With the help of the Service Provider group, we have set up multiple KPs and made our knowledge graphs available to ARAs through both Smart API and TRAPI endpoints. In addition, we have directly provided KGX-formatted TSV files to the (im)Prove Agent and Unsecret Agent ARA groups. As described below, Service Provider tools have helped us streamline the process of deploying and updating KPs, allowing us to focus more time and effort on knowledge generation. Most recently, we have been able to make multiple, incremental updates to our wellness KP (currently on v1.3) and also substantially scale up our EHR Risk KP specifically to support the Workflows B, C, and D and other use cases of interest to Translator. Currently, we are also populating our Big GIM I KP (with gene-gene expression correlations) with ~300 million edges and adding ~107 million new edges (drug-gene expression associations) to our Big GIM II Drug Response KP.

1) The Service Provider group provides separate TRAPI endpoints that allow TRAPI-compliant queries to be submitted to one or more KPs at different scopes. This serves as a proxy API that converts TRAPI queries to Smart API queries, collects responses from KPs and returns a TRAPI response to a calling ARA. TRAPI compliance and CI/CD for the API is handled entirely by the Service Provider group.
2) The */v1/query* endpoint queries and returns results from all registered Smart API KPs hosted on the Service Provider server at https://biothings.ncats.io. The */v1/smartapi/{smartapi_id}/query* endpoint queries and returns results from a single specified Smart API KP. TRAPI endpoints for Multiomics KPs can be found here: https://github.com/NCATSTranslator/Translator-All/wiki/Multiomics-Provider.
3) The Service Provider group has helped us set up several Smart APIs that can be queried directly or through the TRAPI endpoints. KP development and deployment consists of the following steps: (a) Generate knowledge from correlation analysis, machine learning model predictions, etc. and structure it in the form of nodes and edge relationships following the KGX format, (b) store the KGX TSV files on a file server, (c) write a parser script and manifest file as described in the BioThings Studio documentation and store these files in a GitHub repository, and (d) use BioThings Studio to deploy to the Service Provider server. A CI/CD flag can be enabled by the Service Provider group for step (d) so that updated KGX files stored on a file server can easily be loaded and merged with the existing KP.

Parsers (for KP ingest using Service Provider tools), Smart API specifications (for KP registration and defining how TRAPI queries are processed), and other utilities to streamline KP deployment can be found in the following Github repositories:

- Multiomics Wellness KP: https://github.com/Hadlock-Lab/multiomics_wellness_kp
- EHR Risk KP: https://github.com/Hadlock-Lab/clinical_risk_kp
- Big GIM I: https://github.com/Hadlock-Lab/biggim_kp
- Big GIM II: Drug Response KP: https://github.com/Hadlock-Lab/drug_response_kp
- Big GIM II: TCGA Mutation KP: https://github.com/Hadlock-Lab/tcga_mut_freq_kp

4. Integration with other tools and contribution to the Program

Participation in workgroups:

Our team regularly participates in core Translator workgroups (Data Modeling, Architecture).
Biolink, Predicates: Arpita Joshi
User-centered working group: Guangrong Qin, Gustavo Glusman
EPC working group: Guangrong Qin, Gustavo Glusman
Clinical working group: Jennifer Hadlock, Ryan Roper, Qi Wei
Biweekly stand-ups: Ryan Roper, Jennifer Hadlock, Gustavo Glusman
Minihackathons: Ryan Roper, Jennifer Hadlock
Publications committee: chaired by Gustavo Glusman
We regularly work with the Biolink team to standardize edges to support the new KGs being developed, and with the EPC team to make examples of evidence and provenance for KGs extracted from datasets.

Integration with other components:

Multiomics KPs are strongly connected with the im(Prove) Agent, which has two modes of operation: (1) based on the pre-computed SPOKE network, and (2) by building on-the-fly KGs in response to the query for too-large-to-handle knowledge sources. In both modes, Multiomics

KPs offer valuable support to im(Prove) Agent, significantly expanding the scope of connectivity in the BigKG of SPOKE:

1. In a developmental version, SPOKE has ingested co-occurrence information of clinical variables from Multiomics EHR Risk, Clinical Data Provider and from Multiomics Wellness Provider as edges. The connections offered by Multiomics Provider are of particular value because they are derived from Wellness studies that measure plasma proteins and metabolites that are usually not measured, hence providing complementary entities that are not available in clinical EHR data. We analyzed these variables and expressed them as edges that represent statistically significant correlation between pairs of (i) proteins, (ii) Proteins and metabolites, and (iii) metabolites-metabolites. These correlations have been added as edges in the SPOKE network.
2. For on-the-fly KGs, the im(Prove) Agent accesses BigGIM as a knowledge source for gene expression correlations across human tissues (computed from GTEx data) via an internal API.

Our Wellness KG further allows the im(Prove) agent to perform graph analytics on the SPOKE BigKG graph to evaluate graph properties of the KG and determine if “meaningful paths" between nodes are on average shorter than random graph paths. To test this, statistically significant correlations between pairs of or analytes (proteins, metabolites) were mapped onto nodes in the SPOKE BigKG. It was found that indeed nodes representing these correlated variables were connected by a path that was significantly shorter than paths connecting two random nodes (**Figure 6**). This result offers the first empirical evidence that the graph structure of the SPOKE network, which was computationally assembled from diverse biomedical medical databases, preserves meaningful information about mechanistic pathways that traverse various domains, most of them never explicitly mentioned in the literature.

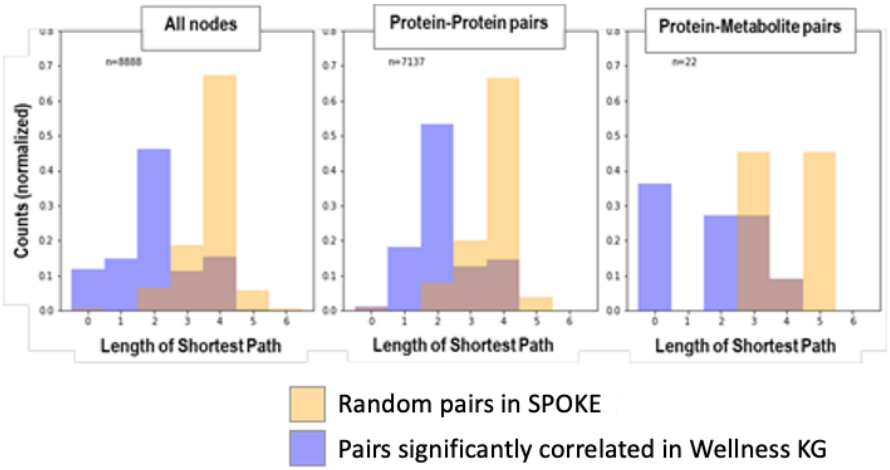


Figure 6: *Analysis of blood proteomics and metabolomic data in healthy participants (Wellness KG) shows that pairs of blood analytes are connected by on average shorter paths in the SPOKE graph than pairs of randomly chosen nodes.*

Furthermore, the Multiomics EHR KP, including the EHR data from PSJH, also helps im(Prove) Agent for learning disease-associated node weights that it uses for ranking the path found using the Google Page Rank algorithm. Here, patients of known diagnosis X (e.g. X=diabetes) “walk randomly” over the SPOKE graph, but in a way that is biased: nodes that represent variables that are altered in the patient with said diagnosis X (e.g. hyperglycemia, neuropath, metformin) are visited more frequently. After learning on a large number of patients, nodes representing variables associated with disease X obtain more weight. This results in disease-specific weight vectors, “PSEV(X)” that im(Prove) Agent uses for empirical data based ranking. The Multiomics

EHR KP enables computing an entire new set of PSEVs that will be invaluable information for the empirical evidence based ranking done by the im(Prove) Agent.

Integration with other Translator components (see also **Figure 7**):
- Exploring Agent (BTE): All Multiomics KPs are integrated via Service Provider
- Expander Agent (ARAX): All Multiomics KPs are integrated in TRAPI 1.1 dev branch
- Ranking Agent (Aragorn): All Multiomics KPs are integrated
- Explanatory Agent: Currently working on integration
- Unsecret Agent: Integrated through KG download
- Clinical Data Services Provider: extensive collaboration on integrating real-world clinical knowledge
- Exposure Provider: extensive collaboration on integrating real-world clinical knowledge

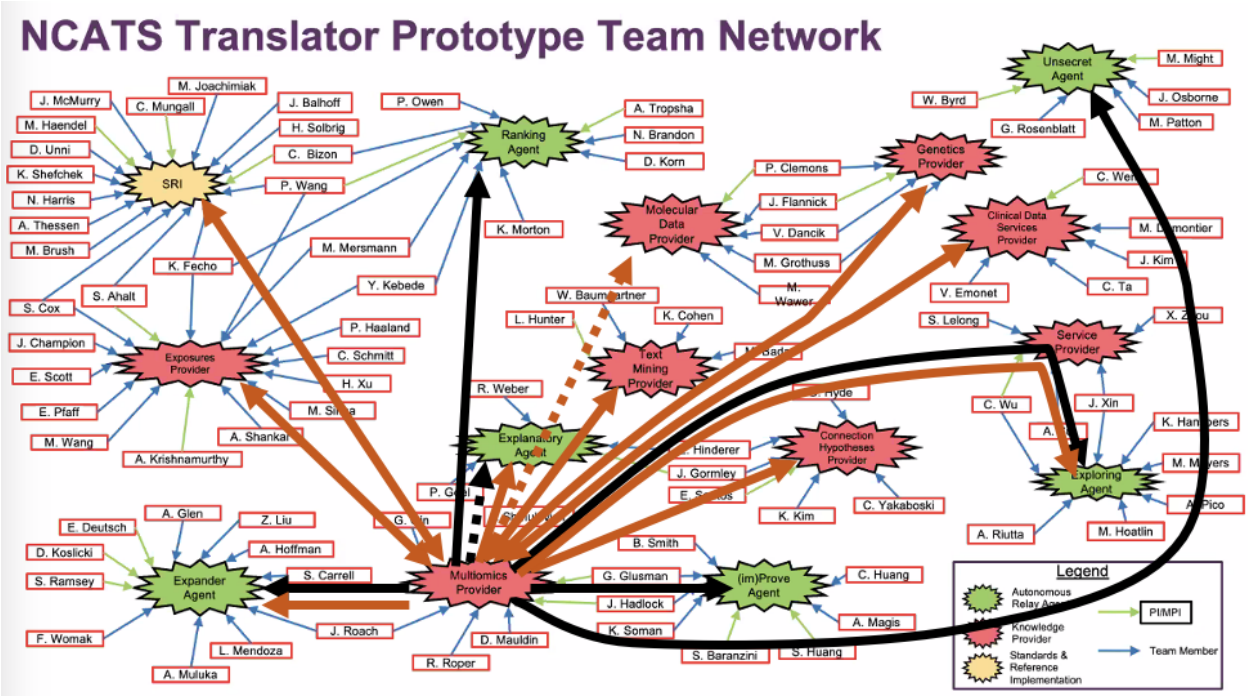


Figure 7: *The Translator Team Network and integration with other components. Black arrows represent KG integration, orange arrows represent collaboration on designing December Demo use cases. Existing connections (full arrows) and in progress (dashed arrows).*

5. Use cases supported

Just Fix It, Stand-up Use Cases and mini-hackathons

The following queries from Just Fix It and Stand-up sessions are directly (without modification) supported or supported with minor modification by Multiomics KPs (Wellness, EHR Risk and Big GIM II).

Directly-supported queries
- Chemical substances associated with dementia
- Chemical substances associated with Alzheimer's
- Chemical substances associated with asthma

- Chemical substances associated with Gene STK11
- Conditions associated with Gene PDZD2

Queries supported with minor modification
- Drugs associated with decreased risk of atrial fibrillation
- Genes associated with breast cancer
- Diseases associated with beclomethasone
- Phenotypes associated with influenza
- Correlation with glycerol

Clinical Data WG Drug-induced Liver Injury Use Case

As part of the Clinical Data WG, we have been coordinating with ARA and other KP groups to begin designing and testing a use case focused on drug-induced liver injury (DILI) that takes advantage of multiple Multiomics and other Translator KPs. Our **Multiomics Wellness KP** provides clinical labs (ALT, AST, ALP) and metabolites and **Multiomics Big GIM II KP** provides genes associated with hepatic function. We substantially scaled up (>20-fold addition of edges) our **Multiomics EHR Risk KP**, in part to be able to support a wide range of use cases, including some rare diseases.. Specifically, we now provide the following associations:

Precision medicine - use cases

A use case raised in the precision medicine section from May 2021 Relay: A patient has a phenotype of developmental delay, speech delay, and seizures, and with RHOBTB2 (Rho GTPase) mutation. Could we find Drug/compound to inhibit RHOBTB2 in the central nervous system? It is a general question that given a patient features or drived by certain alterations. Which drugs could we consider for further estimation or treatment.
The BigGIM II KGs could help to find candidate drugs or targets that could be used for further evaluation. As RHOBTB2 is a candidate tumor suppressor gene, which is not easily to be targeted, there are currently no FDA approved drugs that can target RHOBTB2. We then infer from other diseases (cancers) about how the alteration of the gene is associated with sensitivity to certain drugs or chemicals using **Multiomics BigGIM II - Drug response KP**, and we found that JAK inhibitor Ruxolitinib, MAP2K5 inhibitor XMD8-85 shows higher sensitivity to certain cancer types. Using **Multiomics BigGIM - Gene Gene Interaction KP,** we inferred from the gene gene interactions or regulations, and got the list of genes which show significant correlation with RHOBTB2, and targeted by FDA approved drugs, shown in **Table 2**. It provides new clues for users to explore new possibilities for testing and evaluation.

More use cases can be supported from our user engagement plan. The knowledge graphs have been proved to be useful during the Sep 2020 and May 2021 relays, and we're working closely with ARAs and the ARS to advance integration.

Table 2. *Drug targets that show significant associations with RHOBTB2 in healthy brain samples (data source: GTEx).*

Gene1	Gene2	rho_spearma	pvalue	Tissue type	Is.target
RHOBTB2	THRB	0.84287931	0	Brain	THRB
RHOBTB2	SCN3B	0.79952374	0	Brain	SCN3B
RHOBTB2	SLC8A1	0.75511213	0	Brain	SLC8A1
RHOBTB2	RORB	0.74860527	0	Brain	RORB
RHOBTB2	TACR1	0.73199701	0	Brain	TACR1
RHOBTB2	SLC1A2	0.70143183	0	Brain	SLC1A2
RHOBTB2	SHBG	0.66347242	0	Brain	SHBG
RHOBTB2	TAC3	0.66231144	0	Brain	TAC3
RHOBTB2	SLC6A1	0.65329041	2.243e-321	Brain	SLC6A1
RHOBTB2	TERT	0.64902496	7.6415009e-316	Brain	TERT
RHOBTB2	TNNC2	0.63092155	2.51E-293	Brain	TNNC2
RHOBTB2	SLC1A1	0.62149603	3.46E-282	Brain	SLC1A1
RHOBTB2	SLC15A2	0.61652817	1.82E-276	Brain	SLC15A2
RHOBTB2	TNIK	0.60377454	3.04E-262	Brain	TNIK
RHOBTB2	SLC25A6	0.60090199	3.97E-259	Brain	SLC25A6
RHOBTB2	SLC7A3	0.59498335	8.29E-253	Brain	SLC7A3
RHOBTB2	VDAC1	0.592735	1.93E-250	Brain	VDAC1
RHOBTB2	SCARB1	0.57831113	1.08E-235	Brain	SCARB1
RHOBTB2	SCN5A	0.57475798	3.57E-232	Brain	SCN5A
RHOBTB2	RXRA	0.57146484	6.00E-229	Brain	RXRA
RHOBTB2	SSTR1	0.56580042	1.74E-223	Brain	SSTR1
RHOBTB2	SST	0.56214783	5.09E-220	Brain	SST
RHOBTB2	SLC25A18	0.56116737	4.26E-219	Brain	SLC25A18
RHOBTB2	SRC	0.55356773	4.79E-212	Brain	SRC
RHOBTB2	STK24	0.54962011	1.87E-208	Brain	STK24
RHOBTB2	STK16	0.54500779	2.56E-204	Brain	STK16
RHOBTB2	TOP1MT	0.52817865	9.21E-190	Brain	TOP1MT
RHOBTB2	UTRN	0.50617696	6.26E-172	Brain	UTRN
RHOBTB2	SEC14L2	0.50543087	2.39E-171	Brain	SEC14L2

Drug repurposing for multiple sclerosis and immune-mediated inflammatory diseases

In the April edition of the Translator Gazette, and subsequent follow up use case, we reported on a drug repurposing use case we have been working on for multiple sclerosis (MS). Querying our **Multiomics EHR Risk KP** for medications (biolink:ChemicalSubstance) related to multiple sclerosis, we got the following top results:

subject_id	subject_name	predicate	object_id	object_name	feature_coefficient
UNII:3JB47N2Q2P	natalizumab	biolink:related_to	MONDO:0005301	multiple sclerosis	8.366111
UNII:A10SJL62JY	ocrelizumab	biolink:related_to	MONDO:0005301	multiple sclerosis	8.158063
CHEBI:2972	baclofen	biolink:related_to	MONDO:0005301	multiple sclerosis	3.281209
CHEBI:135738	clevidipine	biolink:related_to	MONDO:0005301	multiple sclerosis	3.247821
CHEBI:2637	amikacin	biolink:related_to	MONDO:0005301	multiple sclerosis	2.567319
CHEBI:68841	gadobutrol	biolink:related_to	MONDO:0005301	multiple sclerosis	2.253569
CHEBI:3403	carboprost	biolink:related_to	MONDO:0005301	multiple sclerosis	2.193086
UNII:4F4X42SYQ6	rituximab	biolink:related_to	MONDO:0005301	multiple sclerosis	2.114505
CHEBI:204928	cefotaxime	biolink:related_to	MONDO:0005301	multiple sclerosis	2.098101
CHEBI:144551	oxybutynin	biolink:related_to	MONDO:0005301	multiple sclerosis	2.01286
CHEBI:63629	tizanidine	biolink:related_to	MONDO:0005301	multiple sclerosis	1.589938
CHEBI:84276	methylphenidate	biolink:related_to	MONDO:0005301	multiple sclerosis	1.497206
CHEBI:15882	phenol	biolink:related_to	MONDO:0005301	multiple sclerosis	1.471017
CHEBI:6888	methylprednisolone	biolink:related_to	MONDO:0005301	multiple sclerosis	1.41143
CHEBI:45783	imatinib	biolink:related_to	MONDO:0005301	multiple sclerosis	1.179723

The query graph itself and links to ARS results can be found here. Imatinib is used for treatment of chronic myelogenous leukemia (CML), but is also currently in clinical trials for repurposing for MS. To discover what other drugs might be repurposed for MS, based on the mechanism of action of imatinib, we ran a two-hop query through the ARS to identify genes associated with both MS and imatinib. Results from ARAX are shown in **Figure. 8**. We then extended this to a three-hop query to identify other drugs associated with this set of genes. We encountered errors from ARAs as reported in github #57. David Koslicki subsequently ran this query directly on ARAX using their DSL and was able to get results, which can be viewed here: https://arax.ncats.io/beta/?r=8461. See **Figure 9**. We are currently reviewing these results. One example of a compelling result (Result 3 in the ARAX results) is nimodipine. According to one study [14]: "*Treatment with nimodipine restores spinal oxygenation and can rapidly improve function. Nimodipine therapy also reduces demyelination in both [experimental autoimmune encephalomyelitis (EAE)] and a model of the early MS lesion*."

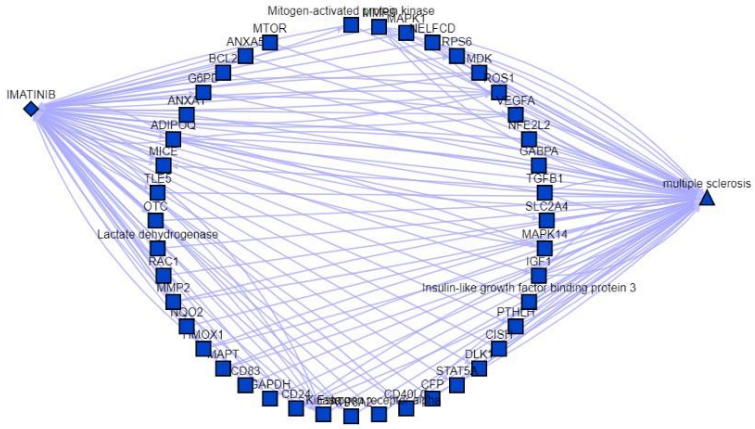


Figure 8: *Drug repurposing for MS, ARAX results.*

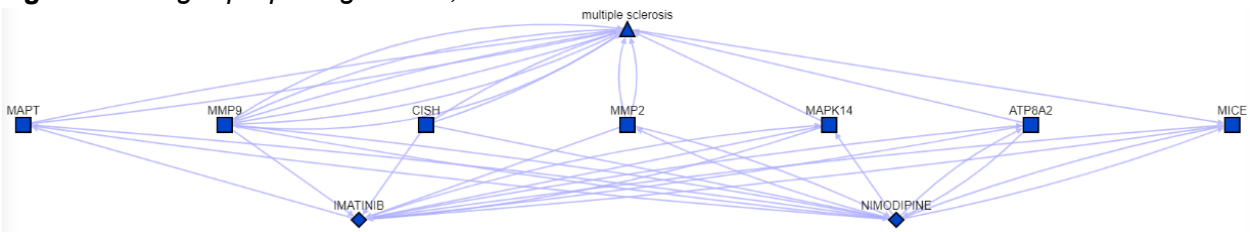


Figure 9: *Example of drug with potential relevance for MS treatment.*

As described in Section 1, we proactively worked to incorporate a wide range of KPs and ARAs and expand this into a full demo with three use cases, as fully described in [github Workflow C](). Some highlights are shown here.

Clinical real-world evidence: current drugs and potential new insights
Immune-mediated inflammatory diseases (IMIDs)

Figure 10: *Demo C highlighting how the need to investigate multiple diseases and drugs, moving beyond the idea of one disease, one target, one drug.*

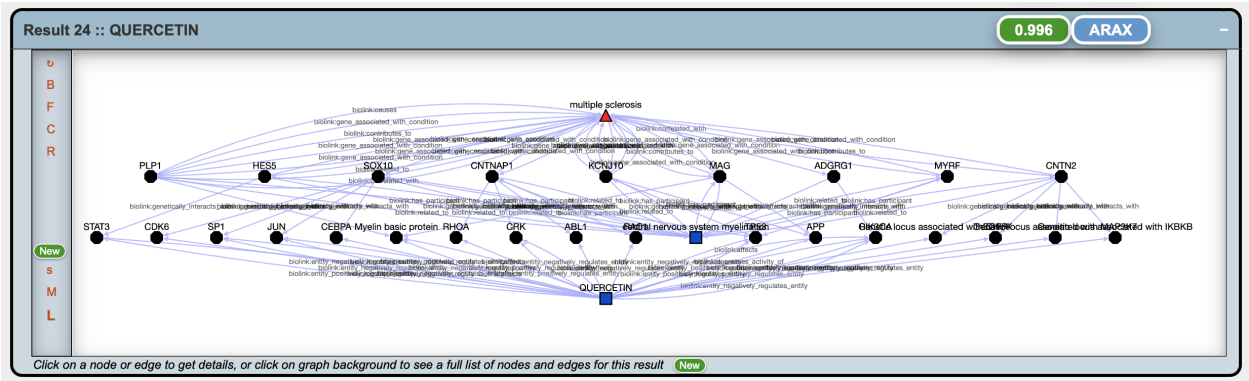


Figure 11: *Example of how Translator adds insight on drugs of known interest to SMEs.*

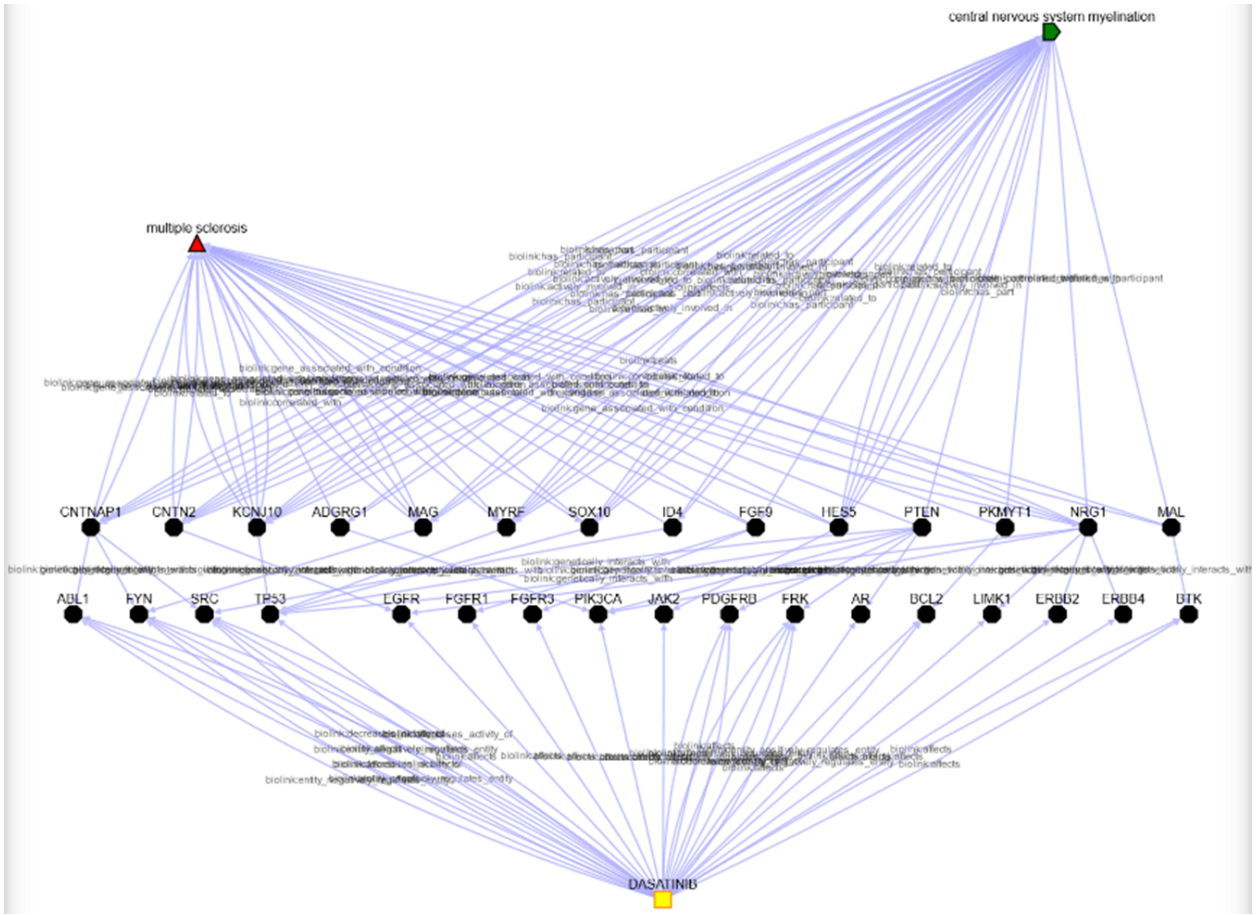


Figure 12: *Example of one of the drugs Translator surfaced that is of interest to SMEs for followup.*

6. Challenges encountered and potential solutions

- Gaps between our KGs and other KGs. From a data driven approach to provide evidence of connection between two entities, the contexts such as disease types behind the data are important. Currently, we are using MONDO ids to label diseases. As there are several MONDO identifiers for the same disease category, such as lung cancer, an exact match might result in no overlap between one KG to the others. A hierarchical annotation for different identifiers might help to solve this problem.
- Similarly, a challenge was mapping concepts to ontologies that can be used (e.g., while LOINC codes are available for our chemistries table, they are not readily usable by other Translator components). Finding a way to convert LOINC codes to equivalent alternate ontology: e.g., three different nodes with different LOINC codes involve zinc, do these nodes get combined into one node in an alternate ontology? What about measurements that don't exist in any other alternate ontology?
- The actual Wellness graph is more complex and diverse than what is served via the KP because of naming difficulties for certain nodes. We expect these to be resolved gradually over time through ongoing curation.
- Identifying, tracing and rationalizing multiple arbitrary encoding decisions that were made while generating and organizing the data, due to real-world complexity.

- Optimization: raw multiomic data can be very large and involve significant computational effort.
- Defining and codifying the context in which measurements were made, to compute their relevance to the context in which the question is being asked.
- One concept can be characterized by different measurements. For example, when we are modeling the COVID-19 severity associated molecular features, there are at least two ways to characterize the COVID-19 severity. The first one is the WHO scale from 1 to 7, and the second is the survival of the patient. To help users distinguish between different models, more precise annotation about the graph will be important.
- During the user engagement process, one researcher was interested in understanding Fanconi anemia, which has an incidence rate of 1 out of 136,000 newborns. It is difficult to generate direct knowledge graphs from a large population of datasets for rare diseases. A promising way to uncover the pathology or therapeutic options will be to borrow the information from other related diseases, such as acute myeloid leukemia which are a common progressed disease for the fanconi anemia patients, and these knowledge graphs can be defined more comprehensively using large data resources which has been included in the developed knowledge graphs. One approach to use the BigGIM II to facilitate the exploration of rare disease (for example, Fanconi anemia) is rationaled as follows, which could further be developed in actionable workflows. Domain expertise from Fanconi anemia has shown this disease is featured with deficiency of DNA damage repair gene. By exploring the BigGIM II KP for the gene-gene interactions and drug response KP using a list of Fanconi anemia associated genes could help us to find the potential abnormality of other genes or find potential effective drugs for the Fanconi anemia associated cancers or even the Fanconi anemia patients.
- The extension of the biolink models. For example, one of our KGs about disease severity associated molecular features is beyond the current biolink annotation. From the KG perspective, the more precisely we annotate the edges, the higher value the KG will contribute. From the ARA or ARS perspective, the exact matching of a specific term or association, the higher value the KG will provide. To fill the gap between the precision of the annotation for each KP and the categorized CURIE the biolink provides, an iterable and dynamic biolink model will be valuable.
- Using a data driven approach, most of the KGs we provide include evidence, provenance and version information. The current biolink model guided KGX format is most emphasized on the nodes, edges, however the presentation of the KGs or usage of the KGs with evidence, provenance is still in active discussion. We are currently working with the EPC working group, trying to figure out a better practice of presenting KGs with evidence and provenance which will also be compatible with the TRAPI standard.
- Representation of interactions knowledge (as in Multiomics Wellness KG) requires an association between three nodes (**Figure 13**). This poses a limitation for the current Biolink model, which expects an association to only exist between two nodes (a subject and an object). This is a specific case of an N-ary relationship (see GitHub issue: 566) for which we are currently working on with the Data Modeling and SRI teams to find a suitable solution.

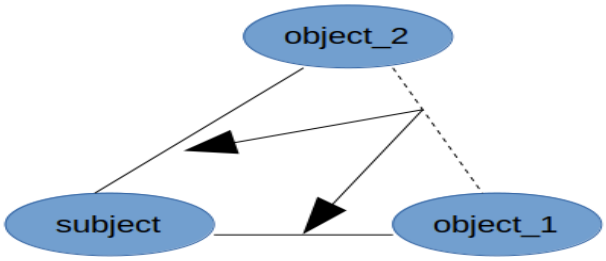


***Figure 13**. Interaction between nodes object_1 and object_2 would affect the relationship (residualized relationship) between subject & object_1 and subject & object_2*

7. Changes from original plans

Ryan Roper has successfully moved on to his goal to work on large scale data engineering. We have welcomed Qi Wei, PhD, a computer scientist who will be supporting Multiomics KPs.

8. User engagement

Current engagement model: In the development stage of our KPs, we developed internal approaches to integrate multiple knowledge graphs we constructed, engaged users to evaluate how valuable the knowledge graphs are, and asked questions which could be answered to address the most challenging and impactful questions. In the production phase, we deployed the knowledge graphs through SmartAPI, queryable through the ARS and ARAX system. We further broadcast the current production version with example queries to the users.
For the development phase, we have been circulating questions in our research community, for example the ISB and collaborators. We started the user engagement by sharing our current capacity as a team as well the capacity as a consortium, then asked if a user would like to share some of the clinical or research questions they would be interested in. For the production phase, we have demonstrated to the users how to use the translator tools to query the KGs we constructed, such as using the ARAX interface.
We met with eight different scientists (seven outside of Translator) from five different institutions and areas of research, and conducted nine user engagement sessions. In each case, we started with broad open ended questions, listening to what challenges each SME was most interested in having Translator help solve. This highlighted some current limitations of Translator that could be addressed in the future, and provided several fruitful use cases to help drive testing and demos. Early on, the work primarily helped reveal new bugs and performance issues across Translator, which we drove to resolution. Over time, as Translator has become more performant, functional and stable, we were able to dive into increasingly complex real-world queries. This work led directly to creating new demos that provided genuinely interesting results for SMEs, and a series of use cases to help prioritize goals for 2022.
Includes a plan for engaging prospective users throughout the development process; roughly one page that includes description of user persona, use-cases, acceptance tests, frequency of interaction with users and the development effort reserved to implement that feedback.
With the knowledge in our BIG GIM II tissue-specific gene gene interactions, we are able to answer specific questions asked by many of the questions related to cancer. For rare diseases, we found it is a big challenge for researchers to get enough datasets or knowledge to understand these types of diseases. To infer the large datasets for other related diseases are considered useful for the investigation of rare diseases.
Current workflows with users:

- Workflow C1, C2, C3, Immune-mediated inflammatory diseases (including many rare): Philip Mease, MD, Rheumatology, University of Washington, Michael Chiorean, MD, Gastroenterology, Swedish Research Institute, and LuLu Iles-Shih, MD, Gastroenterology, Swedish Research Institute.
- Workflow C2, C3, Neurology, central nervous system de- and remyelination,: Sergio Baranzini, PhD, Neurology, University of San Francisco.

Investigating future workflows:

- Genetics, cancer, Fanconi anemia: Ray Monnat, PhD Genome Sciences and Pathology, University of Washington. (Workflow in develop: Exploring Fanconi anemia by borrowing information from cancer and normal gene interactions and drug response)

- Genetics, Precision medicine in cancer: Chris Kemp, PhD, Precision oncologist, Fred Hutch Research Center and Russell Moser, PhD, Biochemistry, Fred Hutch Research Center, Translational scientist. (Workflow in develop: Overcome drug resistance from specific genetic alterations)
- Diamond Blackfan Anemia: Chris Lausted, PhD, ISB, and Raymond Doty, PhD Hematology, University of Washington.

9. Deployment status

- Deployment of our KPs to NCATS cloud services is performed via Service Provider.
- BDEx: a python library that transform data into knowledge (https://github.com/gloriachin/BDEx)
- Parser to transform the knowledge graphs to TRAPI standard format
- Multiomics-KP-API: An endpoint for testing multiomics and querying the multi-omics KGs (http://35.233.133.157:5000/docs)
- Deployed APIs in the SmartAPI endpoints.

10. Unobligated funds

- *Anticipated **unobligated funds** remaining at the end of this reporting period; ensuring to include why funds are remaining, plan for future expenditure and why the funds would be required beyond this budget segment end date.*

7. References

1. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46:D1074–82.

2. Ursu O, Holmes J, Bologa CG, Yang JJ, Mathias SL, Stathias V, et al. DrugCentral 2018: an update. Nucleic Acids Res. 2019;47:D963–70.

3. Wang Y, Zhang S, Li F, Zhou Y, Zhang Y, Wang Z, et al. Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. Nucleic Acids Res. 2020;48:D1031–41.

4. Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, et al. A comprehensive map of molecular drug targets. Nat Rev Drug Discov. 2017;16:19–34.

5. Price ND, Magis AT, Earls JC, Glusman G, Levy R, Lausted C, et al. A wellness study of 108 individuals using personal, dense, dynamic data clouds. Nat Biotechnol. Nature Publishing Group; 2017;35:747–56.

6. Wilmanski T, Rappaport N, Earls JC, Magis AT, Manor O, Lovejoy J, et al. Blood metabolome predicts gut microbiome α-diversity in humans. Nat Biotechnol. Nature Publishing Group; 2019;37:1217–28.

7. Wilmanski T, Diener C, Rappaport N, Patwardhan S, Wiedrick J, Lapidus J, et al. Gut microbiome pattern reflects healthy ageing and predicts survival in humans. Nat Metab. 2021;3:274–86.

8. Alikaşifoğlu A, Gönç N, Özön ZA, Sen Y, Kandemir N. The relationship between serum adiponectin, tumor necrosis factor-alpha, leptin levels and insulin sensitivity in childhood and adolescent obesity: adiponectin is a marker of metabolic syndrome. J Clin Res Pediatr Endocrinol. 2009;1:233–9.

9. Eddy JA, Hood L, Price ND, Geman D. Identifying tightly regulated and variably expressed networks by Differential Rank Conservation (DIRAC). PLoS Comput Biol. 2010;6:e1000792.

10. Joshi A, Haspel N. A Novel Data Instance Reduction Technique using Linear Feature Reduction. AIS. 2020;191–206.

11. Robinson M, Hadlock J, Yu J, Khatamian A, Aravkin AY, Deutsch EW, et al. Fast and simple comparison of semi-structured data, with emphasis on electronic health records [Internet]. bioRxiv. 2018 [cited 2018 Apr 14]. p. 293183. Available from: https://www.biorxiv.org/content/early/2018/04/02/293183

12. Su Y, Chen D, Yuan D, Lausted C, Choi J, Dai CL, et al. Multi-Omics Resolves a Sharp Disease-State Shift between Mild and Moderate COVID-19. Cell. 2020;183:1479–95.e20.

13. Dai CL, Kornilov SA, Roper RT, Cohen-Cline H, Jade K, Smith B, et al. Characteristics and Factors Associated with COVID-19 Infection, Hospitalization, and Mortality Across Race and Ethnicity. Clin Infect Dis [Internet]. 2021; Available from: http://dx.doi.org/10.1093/cid/ciab154

14. Desai RA, Davies AL, Del Rossi N, Tachrount M, Dyson A, Gustavson B, et al. Nimodipine Reduces Dysfunction and Demyelination in Models of Multiple Sclerosis. Ann Neurol. 2020;88:123–36.