bioRxiv preprint doi: https://doi.org/10.1101/2022.01.22.477231; this version posted January 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Subject Section

Context-aware information selection and model recommendation with ACCORDION

Yasmine Ahmed[1, *], Cheryl Telmer[2] and Natasa Miskov-Zivanov[1,3, *]

[1]Electrical and Computer Engineering Department, [3]Bioengineering, Computational and Systems Biology, University of Pittsburgh, [2]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Querying new information from knowledge sources, in general, and published literature, in particular, aims to provide precise and quick answers to questions raised about a system under study. In this paper, we present ACCORDION (**ACC**elerating and **O**ptimizing model **R**ecommen**D**at**ION**s), a novel methodology and a tool to enable efficient addressing of biological questions, by automatically recommending models that recapitulate desired dynamic behavior. Our approach integrates information extraction from literature, clustering, simulation and formal analysis to allow for automated, consistent, and robust assembling, testing and selection of context-specific models. We used ACCORDION in nine benchmark case studies and compared its performance with other previously published tools. ACCORDION is comprehensive as it can capture all relevant knowledge from literature, obtained by automated literature search and machine reading. At the same time, as our results show, ACCORDION is selective, recommending only the most relevant and useful subset (15-20%) of candidate model extensions found in literature, while guided by baseline model context and goal properties. ACCORDION is very effective, also demonstrated by our results, as it can reduce the error of the initial baseline model by more than 80%, recommending models that closely recapitulate desired behavior, and outperforming previously published tools. In this process, ACCORDION can also suggest more than one highly scored model, thus providing alternative solutions to user questions and novel insights for treatment directions.

Contact: {yaa38, nmzivanov}@pitt.edu
Supplementary information: Supplementary data are available.
Availability: http://www.biodesignlab.pitt.edu/accordion

1. Introduction

While modeling helps explain complex systems, guides data collection and generates new challenges and questions [1], it is still largely dependent on manual human contributions. To collect useful information and create reliable models in biology, modelers survey hundreds of papers, search model and interaction databases (e.g., Reactome [2], STRING [3], KEGG [4], etc.), incorporate background and common-sense knowledge of domain experts, and interpret results of wet-lab experiments. These time-consuming steps make the creation and the development of models a slow, laborious and error-prone process.

On the other hand, machine learning and bioinformatics advances have enabled automated inference of models from data. Although very proficient in identifying correlations between system components, these methods still struggle if tasked with finding directionality of influences and causation [5][6]. Inferring large causal models from data requires significant time and computational resources, and it is strongly dependent on the quality of the data [7]. Moreover, as the amount of biological data in the public domain grows rapidly, problems of data inconsistency and fragmentation are arising [8].

To overcome issues with data reliability, automating what used to be a manual process in model creation seems like a critical next step in computational modeling. In other words, automated selection of new, reliable, and useful information about component influences and causality, followed by recommendation of how to add them to models will be beneficial in several ways. Besides leading to more *efficient* modeling by removing slow manual steps, it will allow for more *consistent*, *comprehensive*, *robust*, and better *curated* modeling process.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.22.477231; this version posted January 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In order to automate the collection of relevant and useful information about component influences and causality, one can begin with a query about the system, its components, behavior, or features of interest. The search query guides automated selection of articles that contain relevant information from published literature databases. The biomedical literature mining tools are essential for the high throughput extraction of knowledge from scientific papers, examples of such reading engines are REACH [9], TRIPS [10], Eidos [11]. INDRA (Integrated Network and Dynamical Reasoning Assembler) [12] is an automated model assembly tool designed for biomolecular signaling pathway models and generalized to other domains such as disease models. INDRA relies on collecting and scoring new information extracted either from the textual evidence found in the corpus using an ensemble of natural language processing techniques (including REACH, TRIPS, and Eidos) or from structured pathway databases such as SIGNOR [13]. To select the most valuable and high-quality statements, INDRA computes an overall belief score for each statement which is defined as the joint probability of correctness implied by the evidence.

Recently, several methods have been proposed to automate the process of model extension from the information in literature [14][15]. In [16], the authors describe a method that, in the context of a given baseline model, automatically selects a subset of element interactions from a large machine reading output. The goal of the work described in [16] is to build a model that satisfies a set of requirements or to identify new therapeutic targets, formally expressed as existing or desired system properties. The main drawback of the method in [16] is that it becomes impractical for large models due to adding new interactions in layers, based on their proximity to the existing model. Another model extension method was proposed in [17], it uses a genetic algorithm and it is able to select a set of extensions from machine reading output that lead to a new model with desired behavior. The two main disadvantages of this approach are issues with scalability and the non-determinism, as the solution may vary across multiple algorithm executions on the same inputs. Finally, in [18], the authors proposed a tool which uses several metrics that rely on interaction occurrences and co-occurrences in published literature, and accounts for the connectivity of the newly added interactions to the existing models. While it selects new high-confidence interactions well supported by published literature and connected to the baseline model, the tool described in [18] focuses on the static underlying network of a model and does not consider its dynamic behavior.

In this work, we propose ACCORDION (**ACC**elerating and **O**ptimizing model **R**ecommen**D**at**ION**s), a tool that identifies useful and relevant information from published literature, and recommends model modifications that lead to closely recapitulating desired system behavior, all in a fully automated manner. Thus, compared to the work in [18], ACCORDION also considers the dynamic behavior, and in contrast to [16] and [17], it focuses on identifying clusters of strongly connected elements in the newly extracted information, that can have a measurable impact on the dynamic behavior of the model.

ACCORDION is versatile, it can be used to extend any model that has a directed graph as an underlying structure, and update functions for elements, allowing studies of system dynamics (also known as executable models). To demonstrate the efficiency and utility of the tool, we have selected nine different case studies using models of three systems, namely, the T cell differentiation model [19], the T cell large granular lymphocyte model [20] and the pancreatic cancer cell model [21], and seven machine reading outputs with varying features. Our main goal in this work is to show that our tool, ACCORDION, automatically, without human intervention, recommends model improvements to significantly reduce baseline model error and recapitulate desired system behavior.

To this end, the contributions of this work include: (i) a tool that recommends executable models of intracellular signaling to satisfy desired system properties; (ii) a novel approach for integration of machine reading, simulation, and testing that allows for in-design model validation (instead of typical post-design approach); (iii) several new candidate models of the three systems under study, assembled automatically, satisfying the same set of desired properties as existing manually built models, and thus, enabling exploration of redundancies or discovering alternative pathways of regulation. Finally, ACCORDION takes at most a few hours to execute thousands of experiments *in silico*, which would take days, or months, or would be impractical to conduct *in vivo* or *in vitro*.

2. Methods

The inputs and outputs of ACCORDION, as well as the main methods within the tool are outlined in **Figure 1**.

2.1. Baseline model

One of the inputs to ACCORDION is a baseline model (BM), setting the context for other inputs and for the analysis. The baseline model can be obtained in many different ways, for example, it could be manually created with expert input, or adopted from models published in literature [22][20][21][23] and in model databases [4][2][24]. In general, ACCORDION works with models that have *directed cyclic* graph structure, $G(V,E)$, where each node $v \in V$ corresponds to one model element, representing a protein, gene, chemical, or a biological process, and each directed edge $e(v_i, v_j) \in E$ indicates that element v_j is regulated or influenced, directly or indirectly, by element v_i.

We refer to the set of regulators of an element as its *influence set*, distinguishing between positive and negative regulators. ACCORDION assigns to each element v a discrete variable x representing the element's state, such as a level of its activity or amount. Each model element may have a state transition function, referred to as *element update rule*, which defines its state changes given the states of its regulators, thus enabling the study of system dynamics. While the types of elements and their update rules (see Sections 2.3-2.7) are not constrained by the main methods implemented within ACCORDION, they are largely affected by the information that is available in new events (see Section 2.2) and in the baseline model. Most often, the events described in literature are qualitative, for example, only two element states (e.g., inactive/active, absent/present) may be distinguished or relevant, or only two or three levels of concentration may be considered (e.g., low/high or low/medium/high). Causal or Boolean types of regulations and update rules are most suitable in such cases and ACCORDION is also compatible with such qualitative information. The details of model representation and formats accepted by ACCORDION are provided in Section 1S (supplement).

2.2. Candidate event set

Another input to ACCORDION is a set of candidate events (CEs), which can be collected from different sources and created manually or automatically. Since the machine reading of published literature results in large event sets, and therefore, allows for a high throughput processing of available information, we will assume here such automated pipeline, including both machine readers (e.g., the ones described in [9][10]) and INDRA database of interactions extracted from literature [12]. The set of relevant papers can be selected either using search tools such as Google or PubMed [25] or by providing key search terms to reading engines, which then directly use Medline search tools (e.g., PubMed [25], Ovid [26]) to find most relevant papers. The former approach includes manual user step (using search tools to find papers to input to machine readers), but gives

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.22.477231; this version posted January 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

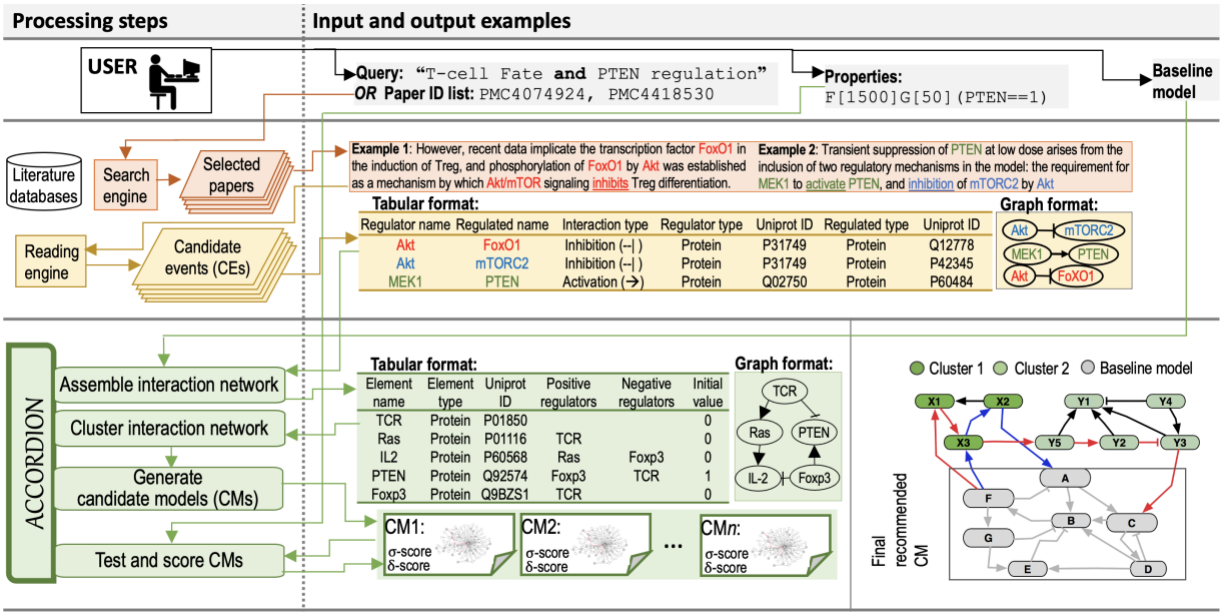


Figure 1. The diagram of the flow that includes a user, machine reading and ACCORDION. Input and output examples column: **(Top)** *Left*: Example query used to select relevant papers. *Right*: example property written in BLTL format. **(Middle)** Main components of information extraction from relevant papers. *Top*: Two example sentences with highlighted entities and events that are extracted by machine readers. *Bottom-Left*: Tabular outputs from REACH engine with **Example 1** and **Example 2** sentences as input. *Bottom-Right*: Graphical representation of REACH outputs. **(Bottom)** *Left*: Tabular representation of several elements and their influence sets (positive and negative regulators) in BioRECIPES format [45] and the graphical representation of elements and influence sets. *Right*: A toy example graph G^{new} of a baseline model and connected clusters: grey nodes belong to the baseline model, light and dark green nodes belong to the CE set obtained from machine reading, blue edges highlight a return path within one cluster, and red edges show a return path connecting two clusters. The multi-cluster path starts at G^{BM} (baseline model), continues through C_1 (cluster 1), then through C_2 (cluster 2), and ends in G^{BM}.

more flexibility to users when selecting relevant papers, while the latter approach allows for full automation, starting with the query entered by a user. In either case, machine readers process the selected papers and output a set of events. Examples of queries, sentences processed by machine readers, and events in the machine reading output are shown in **Figure 1**. As can be seen in the figure, each event has a direction (source and target of interaction) and sign (positive or negative regulation).

2.3. G^{new} creation and return path definition

The CE set can be represented as a set of edges, E^{ext}, where the source and target nodes of these edges form set V^{ext}. From the baseline model graph $G^{BM}(V^{BM}, E^{BM})$ and the CE set, ACCORDION creates a new graph $G^{new}(V^{new}, E^{new})$, where $V^{new} = V^{BM} \cup V^{ext}$, and $E^{new} = E^{BM} \cup E^{ext}$. The edges $e(v_s,v_t)$ in E^{ext}, where v_s is the source node and v_t is the target node, can be classified into three categories: (i) both source node v_s and target node v_t are found in the baseline model: $\{v_s,v_t\} \in V^{BM}$; (ii) either the source node or the target node is found in the baseline model: ($v_s \in V^{BM}$ and $v_t \notin V^{BM}$) or ($v_s \notin V^{BM}$ and $v_t \in V^{BM}$); (iii) neither the source node nor the target node is found in the baseline model: $\{v_s,v_t\} \notin V^{BM}$.

Adding the entire set of CEs to the baseline model all at once usually does not result in a useful and accurate model. Alternatively, we can add one interaction at a time and test each model version, which is time consuming, or even impractical, given that the number of models increases exponentially with the size of the CE set. Moreover, adding individual interactions does not have an effect on the model when an interaction belongs to category (iii), and most often when it belongs to category (ii). It proves much more useful to add paths of connected interactions, which are at the same time connected to the baseline model in their first and last nodes. Therefore, our approach for finding the most useful subset of the CE set includes finding connected interactions, that is, a set of edges in the graph G^{new} that form a *return path*. We define a path of k connected edges as $e^{path}(v_{s1},v_{tk}) = (e_{i1}(v_{s1},v_{t1}), e_{i2}(v_{s2}=v_{t1},v_{t2}), e_{i3}(v_{s3}=v_{t2},v_{t3}), \ldots, e_{ik}(v_{sk}=v_{tk-1},v_{tk}))$, and we will refer to $e^{path}(v_{s1},v_{tk})$ as a return path, when $\{v_{s1},v_{tk}\} \in V^{BM}$ (**Figure 1**). ACCORDION searches for such return paths after clustering G^{new}.

2.4. G^{new} clustering

To find clusters in G^{new}, we apply Markov Clustering algorithm (MCL) [27], an unsupervised graph clustering algorithm, commonly used in bioinformatics (e.g., clustering of protein-protein interaction networks [28][29]). In [30], the authors showed that the MCL algorithm is tolerant to noise, while identifying meaningful clusters. A number of previous studies have demonstrated that the MCL algorithm outperforms other clustering techniques [28][31][32][33][34]. The MCL algorithm has been proven to converge with undirected graphs [30], and therefore, ACCORDION provides to the MCL algorithm the information about node adjacency in G^{new}. Since we are interested in clustering a graph given its connectivity only, the information about adjacency without directionality is sufficient in this step. The directionality will be used in later steps when exploring dynamic behavior. In other words, the adjacency matrix M created this way is symmetric, mapping nodes in G^{new} to both row and column headers in M. The entries in matrix M are assigned value 1 when an edge between their column and row nodes exists in G^{new} or when an entry is on the main diagonal of M (i.e., same column and row node), and value 0 otherwise. Next, the updated matrix M is used by the MCL algorithm as an initial version of a stochastic Markov matrix [35], where each entry represents the probability of a transition from the column node to the row node. Since G^{new} is not a weighted graph, all transitions are

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.22.477231; this version posted January 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

assumed to be equally likely, and the matrix M is normalized such that the sum of entries in each column is equal 1. As mentioned earlier, graph G^{new} can be cyclic, and although the MCL algorithm has been previously applied to acyclic graphs [36], we still use the MCL algorithm for its speed, and our results show that it provides useful results when applied in automated model extension recommendation.

MCL simulates random walks on an underlying interaction network (in our case, graph G^{new}), by alternating two operations, expansion and inflation. The probability of a random walk of length l between any two nodes can be calculated by raising the matrix M to the exponent l, a process called expansion. As the number of paths is likely larger between nodes within the same cluster than between nodes across different clusters, the transition probabilities between nodes in the same cluster will typically be higher in a newly obtained expanded matrix. MCL further amplifies this effect by computing entry-wise exponents of the expanded matrix, a process called inflation [27], which raises each element of the matrix to the power r. Clusters are determined by alternating expansion and inflation, until the graph is partitioned into subsets such that there are no paths between these subsets. The final number of generated clusters, $C_1,\ldots,C_n$, depends on the selected inflation parameter r [27].

As discussed above, ACCORDION clusters the entire G^{new} in order to account for the connectivity with the baseline model, and thus, it likely assigns parts of the baseline model to different clusters. Once the clusters are generated, since we are interested in adding the components of the CE set from the clusters to the entire baseline model, we will refer to the CE (BM) part of a generated cluster l as C_l^{CE} (C_l^{BM}) and to the nodes and edges in such cluster subsets as $V^{C_l,CE}$ ($V^{C_l,BM}$) and $E^{C_l,CE}$ ($E^{C_l,BM}$), respectively.

2.5. Assembly of candidate model networks

From the generated clusters and the baseline model, ACCORDION assembles multiple *candidate models* (*CMs*) as follows. ACCORDION can add clusters one at a time, or in groups. The more clusters or cluster groups are generated, the number of possible cluster combinations grows, and consequently, ACCORDION needs to assemble and test more models. In addition to that, in most cases V^{BM} is smaller than V^{ext}, and E^{BM} is smaller than E^{ext}, and thus, the number of new nodes and edges in a cluster tends to be relatively large compared to the size of the baseline model (we will show examples for our case studies later in Section 3.1). Adding a large number of new nodes and edges to the baseline model at once can significantly change the structure and the behavior of the model. Therefore, the default approach in ACCORDION is to evaluate only individual clusters generated as described in Section 2.3, as well as clusters $C_{i,j}$, created by merging pairs of clusters C_i and C_j ($i,j = 1..n$, $i \neq j$). ACCORDION determines for each individual and merged cluster whether it forms a return path with the baseline model, and for each such cluster, ACCORDION creates a candidate model by adding the entire baseline model to the cluster. In other words, the number of created candidate models is equal the number of clusters (both individual and merged) that form a return path with the baseline model.

As defined above, the clusters formed from the G^{new} graph can contain nodes and edges of the baseline model. Therefore, for those clusters (individual or merged) that were used to create candidate models, ACCORDION computes the *node overlap* (*NO*) value [18], as a ratio of those nodes in a cluster C_l that are present in the baseline model ($V^{C_l,BM} = V^{BM} \cap V^{C_l}$) and the total number of nodes within a cluster (V^{C_l}).

$$NO_l = \frac{|V^{C_l,BM}|}{|V^{C_l}|}$$

2.6. Executable model creation and testing

In previous sections we mostly focused on the static graph structure of the two inputs, baseline model and the CE set. Here, we discuss an additional input to ACCORDION and how all three inputs are used to evaluate the dynamics of candidate models.

The third input to ACCORDION includes a set of properties $\mathcal{T}$ defining desired dynamic behavior that the assembled model should satisfy. ACCORDION uses element update rules in the baseline model and the sign of influences (positive or negative) in the CE set to create new element update rules. For those elements that were already in the baseline model, but their influence set was extended after adding a cluster to the baseline model, ACCORDION modifies their update rules. When new elements with non-empty influence set are added to the baseline model, ACCORDION generates a new update rule for them. As stated previously, event information available in the CE set is often qualitative, for example, "*A* positively regulates *B*". Furthermore, if an update rule for element *B* in the baseline model already includes two positive regulators *C* and *D*, i.e., $x_B = f(x_C, x_D)$, then the new event from the CE set can be added to the update rule for B as $x_B = f(x_C, x_D)$ OR x_A, or $x_B = f(x_C, x_D)$ AND x_A (following the definition from Section 2.1, x_A, x_B, x_C, x_D are variables representing level or amount or activity of elements *A*, *B*, *C*, *D*, respectively). For elements with more than two discrete levels, ACCORDION can use max and min operators to determine the maximum or minimum influence from a given set of regulators.

To select the CM that allows for most closely reproducing the experimentally observed or desired behaviors and, given the randomness in time and order of events in modeled systems, ACCORDION uses a combination of stochastic simulation and statistical model checking. The DiSH simulator, described in detail in [37][38], is used to obtain the dynamic behavior of the baseline model and the CMs. DiSH is a stochastic simulator that can simulate models at different levels of abstraction, information resolution, and uncertainty. This range of simulation schemes is especially valuable when working with diverse information sources and inputs, such as the ones used by ACCORDION. Each simulation run starts with a specified initial model state, where initial values are assigned to all model elements to represent a particular system state (e.g., naïve or not differentiated cell, healthy cancer cell). The initial values for the baseline model elements (nodes in V^{BM}) are usually already known, however, the newly added elements (nodes in V^{ext}) need to be assigned initial values as well. Given that machine reading does not provide this information, we assume that all elements within the same cluster have the same initial value.

ACCORDION runs a statistical model checker [39][40] to verify whether the CMs satisfy a set of properties describing expected behavior of the modeled system. The model checker reads properties formally written using Bounded Linear Temporal Logic (BLTL) [41][40] and, for a given model $\mathcal{M}$ and a property t, it outputs a *property probability estimate*, $p_t^{\mathcal{M}}$, that model $\mathcal{M}$ satisfies property t, under predefined error interval for the estimate. For instance, we can test whether at any point within the first s_1 time steps, model element v_i (i.e., its state variable x_i) reaches value X_1 and element v_j (i.e., its state variable x_j) reaches value X_2, and they both keep those values for at least s_2 time steps. We write this property formally as $F^{s_1}G^{s_2}(x_i = X_1 \wedge x_j = X_2)$, where F^{s_1} stands for "any time in the future s_1 steps", and G^{s_2} stands for "globally for s_2 steps". An example of a property and its expected value are shown in **Figure 1**. To avoid a full state space search, the statistical model checker calls the simulator to generate element trajectories for a defined number of steps and then performs statistical analysis on those trajectories with respect to a given property [42][16].

2.7. CM scoring and recommendation

Usually, we are interested in a model that can satisfy a property $t_j \in \mathcal{T}$ with high probability. However, in some cases, due to randomness in

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.22.477231; this version posted January 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

STUDY	CE set creation								BM creation				BM and CE set relationship		GM		GM and CE set relationship		
	query	pre-selected papers	curation	fixed	read papers	$\lvert E^{RO}\rvert$	$\lvert V^{CE}\rvert$	$\lvert E^{CE}\rvert$	target element	fixed	$\lvert V^{BM}\rvert$	$\lvert E^{BM}\rvert$	$\lvert V^{BM} \cap V^{CE}\rvert$	$\lvert E^{BM} \cap E^{CE}\rvert$	$\lvert V^{GM}\rvert$	$\lvert E^{GM}\rvert$	$\lvert E^{GM} \cap E^{CE}\rvert$	$\lvert E^{GM} \setminus E^{BM}\rvert$	$\lvert (E^{GM} \setminus E^{BM}) \cap E^{CE}\rvert$
Tcell CE^{FA}	✓	-	-	-	11	300	156	239	-		39	60	18	0	43	74	10	14	10
Tcell CE^{SA}	-	✓	✓	-	13	188	82	131	-	✓	39	60	20	2	43	74	13	14	11
Tcell CE^{SM}	-	✓	✓	-	12	118	64	85	-		39	60	15	0	43	74	11	14	11
T-LGL Q^{Sm}	✓	-	-	-	22	78	58	52	-		41	113	16	3	60	193	11	80	8
T-LGL Q^{Med}	✓	-	-	-	38	645	303	448	-	✓	41	113	20	6	60	193	20	80	14
T-LGL Q^{Det}	✓	-	-	-	46	907	424	644	-		41	113	26	5	60	193	18	80	13
PCC BM^{Au}	-	✓	-		19	706	451	631	✓	-	232	323	74	8	257	373	12	50	4
PCC BM^{Ap}	-	✓	-	✓	19	706	451	631	✓	-	195	314	73	6	257	373	12	59	6
PCC BM^{Pr}	-	✓	-		19	706	451	631	✓	-	177	286	73	8	257	373	12	87	4

Figure 2. (Top) Description of each use case in terms of **1-**how each CE set is acquired (using a query or a preselected set of papers, how many papers are read, the number of edges in the entire machine reading output file, $\boldsymbol{E^{RO}}$, the number of nodes or entities $\boldsymbol{V}^{CE}$, the unique number of interactions in each CE set $\boldsymbol{E}^{CE}$), **2-**how each baseline model (BM) is created, whether it is fixed across all three studies for the same system, or it is different for the three studies of the same system, **3-**the relationship between each BM and the corresponding CE set in terms of the number of common nodes (entities) and edges (interactions), **4-**the golden model (GM) specifications, **5-**the relationship between each GM and the corresponding CE set (the number of common edges, the number of edges that are in GM but not in BM, the number of edges that are in GM but not in BM and are found in the CE). **(Bottom)** Venn diagrams showing the overlap between three sets, $\boldsymbol{E}^{CE}$, $\boldsymbol{E}^{BM}$ and $\boldsymbol{E}^{GM}$ for the nine case studies.

biological systems, the $p_t^{\mathcal{M}}$ value lower than 1 (e.g., $p_t^{\mathcal{M}} \geq 0.7$) is expected. In our case studies explored in Section 3 (and in Table 1S), we will show examples of such properties. In order to provide the recommendation of top CMs that are closest to expected probability values for properties, we use several metrics. The first metric, *model property error*, determines the difference between an estimated probability value for property t_j for CM_i, $p_{t_j}^{CM_i}$, and the *goal property probability* value for t_j, P_{t_j}: $\varepsilon_{t_j}^{CM_i} = \left|p_{t_j}^{CM_i} - P_{t_j}\right|$. Next, we compute *average model error*, across all tested properties $t_j \in \mathcal{T}$, for each CM_i, $\varepsilon_{\mathcal{T},avg}^{CM_i}$, and *σ-score* for model CM_i for the given set of properties as $\sigma_{\mathcal{T}}^{CM_i} = 1 - \varepsilon_{\mathcal{T},avg}^{CM_i}$. The larger σ-score for a model is, the closer the model is to satisfying all desired properties. We also define model *δ-score*, $N_{\mathcal{T},\delta}^{CM_i}$, as the percent of properties out of all properties in $\mathcal{T}$ for which $\varepsilon_{t_j}^{CM_i} \leq \delta$. In other words, the parameter δ indicates how close the $p_{t_j}^{CM_i}$ value needs to be to the goal probability P_{t_j} for the property to be considered satisfied. This parameter can be selected by ACCORDION users depending on their modeling goals.

3. Results

3.1. Benchmarks

In the absence of standardized benchmarks to evaluate ACCORDION, we created nine case studies. These benchmarks and all related files will be open access and available with ACCORDION release [43]. In Section 2S in the supplement, we provide an overview of the biological background for all studied systems, the details of creating the baseline model, and the steps of selecting literature and creating CE set for each conducted case study. In Figure 2, we list the main characteristics of these nine cases, with models of three biological systems and different sets of CEs for each system. The three models include control circuitry of naïve T cell differentiation (T cell) [22], T cell large granular lymphocyte (T-LGL) leukemia model [20], and pancreatic cancer cell model (PCC) [21]. The studies vary in the size and graph features of baseline models ("BM creation" columns) and the CE sets (CE set creation" columns), and are named Tcell CE^{FA}, Tcell CE^{SA}, Tcell CE^{SM}, T-LGL Q^{Sm}, T-LGL Q^{Med}, T-LGL Q^{Det}, PCC BM^{Au}, PCC BM^{Ap}, and PCC BM^{Pr}. As can be seen in Figure 2, the size of baseline models varies from several tens to several hundreds of nodes or edges, and the number of interactions in the CE set varies from half the number of interactions in the baseline model to six times larger ("BM and CE set relationship" columns).

We also list in Table 1S (supplement) the sets of desired properties, that are not fully satisfied by baseline models and are used to guide new model assembly for each case study. The properties in Table 1S are provided in both natural language descriptions and machine readable BLTL format, and we also include their goal probability values (P_{t_j}). For each system, besides a baseline model, we also found a golden model in literature ([19] for the T cell model, [20] for the T-LGL model, and [21] for the PCC model). Figure 2 includes the characteristics of golden models (columns "GM" and "GM and CE set relationship").

With these nine case studies, we evaluate ACCORDION's performance and also demonstrate different research scenarios where it can be used, such as varying size and contents of baseline model and CE set (all nine case studies), varying quality of the CE set (Tcell case studies), varying level of detail in user selection of literature (Tcell CE^{FA} and all three T-LGL case studies) reconstruction of previously published model (all nine case studies).

3.2. Recommending new models with desired behavior

In **Figure 3**(a), we show the minimum and maximum of the average model error ($\varepsilon_{\mathcal{T},avg}^{CEM_i}$) found across all created CMs for each of the nine use cases. Additionally, in **Figure 3**(b), we show the δ-score, $N_{\mathcal{T},\delta}^{CEM_i}$, values for the top CMs recommended by ACCORDION in all nine use cases. We also explored different δ values (0.1 to 0.5). To highlight the improvements in CMs when compared to the original baseline model, we show all results next to their corresponding baseline model values. As can be seen from the figure, ACCORDION achieved δ-score of 95% when $\delta = 0.3$ (i.e., all but one property satisfied). Furthermore, increasing δ improves the model score, however, we observed that 0.2 or 0.3 value for δ is optimal to obtain useful models with high score. Overall, ACCORDION automatically selected a small fraction (e.g., ~20%, as will be discussed in Section 3.3) of all interactions in the CE set, sufficient to decrease model error by up to 83%, as shown in **Figure 3**(c).

Furthermore, we compared ACCORDION's performance in terms of average model error of the top recommended model $\varepsilon_{\mathcal{T},avg}^{CMrecommended}$ with two other previously published methods for model extension from [16] and [18]. **Figure 3**(d) shows that ACCORDION obtains the lowest $\varepsilon_{\mathcal{T},avg}^{CMrecommended}$. We applied the layered approach from [16] only on the T cell case study, since it has been shown to mainly work on smaller models, and we applied the approach from [18] on all three baseline

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.22.477231; this version posted January 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

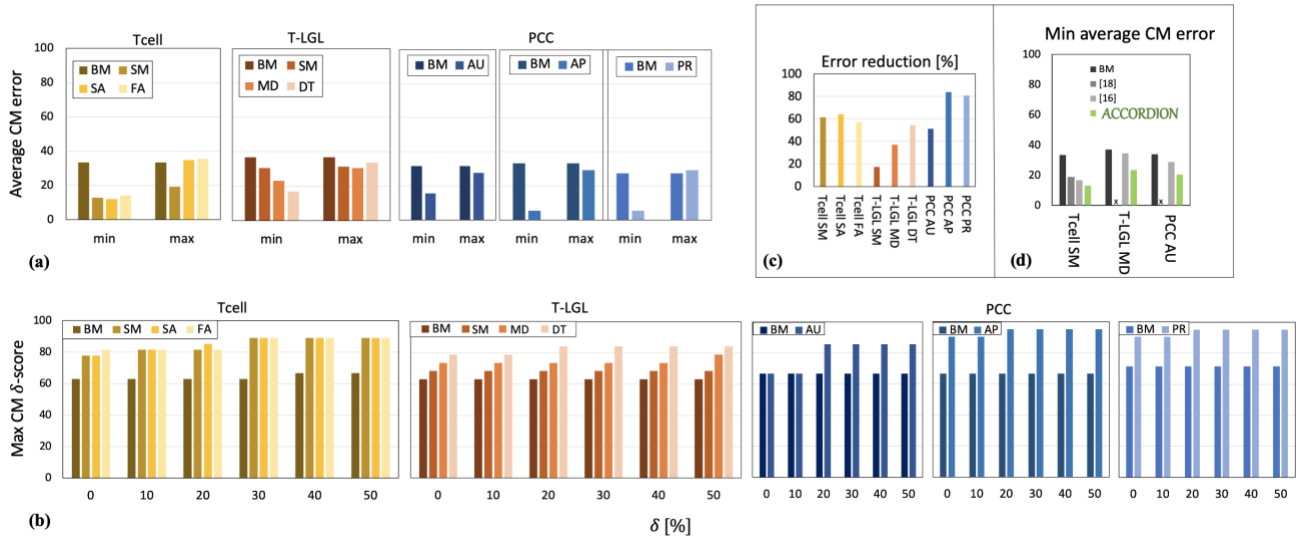


Figure 3. ACCORDION evaluation on nine case studies, three Tcell studies SM, SA, FA, three T-LGL studies SM (Sm), MD (Med), DT (Det), and PCC studies AU, AP, PR: (a) minimum (best) and maximum (worst) average model error $\varepsilon_{\mathcal{T},avg}^{CM_i}$ across all recommended models for each case study, compared to the average error of the baseline model in each study; (b) maximum across all CM δ-scores $N_{\mathcal{T},\delta}^{CM_i}$ obtained in each case study expressed in %; the results are compared for different values of δ (10%, 20%, 30%, 40%, 50%); (c) error reduction ACCORDION achieves in each case study; (d) the comparison between BM error and the top recommended model by ACCORDION and other previously published methods in [16] and [18].

models. The method in [18] relies only on the event occurrences and co-occurrences in literature, without accounting for dynamic behavior, and therefore, ACCORDION outperforms it, as it is guided by the desired system behavior (i.e., the set of properties $\mathcal{T}$ and their corresponding goal property probabilities P_{t_j}).

As can be concluded from **Figure 3**, automated reading and model assembly are not able to reduce model errors $\varepsilon_{\mathcal{T},avg}^{CM_i}$ all the way to 0 in our use cases. ACCORDION outputs $p_{t_j}^{CM_i}$ values for all properties and all CMs it creates, and the list of extensions from CEs that are used in each CM. We show in Figure 1S in the supplement the $p_{t_j}^{CM_i}$ heatmaps that ACCORDION computed for all nine case studies. The heatmaps provide details per each individual property and CM, and this information can be especially useful if users decide to manually inspect and further modify CMs recommended by ACCORDION.

Although we show in Figure 1S results for all properties, several of the CE sets did not fulfill the necessary requirement for all properties to be used. In other words, all the elements that are listed in properties (Table 1S, supplement) need to be present in at least one of the sets V^{BM} and V^{CE}. As shown in **Figure 4** ("Properties" columns), in six out of nine studies, these elements are either already in the baseline model or in the CE set. However, in all three T-LGL studies element GAP is not found in either of the two sets, V^{BM} and V^{CE}, and in the T-LGL Q^{Sm} case two elements, Ceramide and SOCS, are also not present. These element omissions occur in ACCORDION's input and are due to machine reading not finding those elements in selected papers. While the properties that correspond to such omitted elements are not suitable for evaluating ACCORDION, we included them in our results to demonstrate realistic cases with imperfect CE sets. As part of our future work on ACCORDION, we plan to include pre-processing methods to automatically exclude such tests before clustering the CE set, or to inform the user at the beginning that property elements are not found in the input. On the other hand, we were especially interested in ACCORDION's performance in the cases where property elements are not present in V^{BM} but are in V^{CE}. Thus, we defined "criterion A" (Figure 4) to evaluate ACCORDION in such cases. As can be seen from the figure, ACCORDION is able to recover all property elements missing from a baseline model in at least one of the recommended CMs.

Finally, when ACCORDION recovers all necessary property elements, most often the reason for non-zero model property errors ($\varepsilon_{t_j}^{CM_i} > 0$) is in update rules. For instance, in the Tcell cases, for the best recommended model per case, ACCORDION was able to recover FOXO1 which was not in V^{BM} but was in V^{CE}. Moreover, ACCORDION recovered the update function of FOXO1 in all three cases and therefore, the properties that correspond to the dynamic behavior of FOXO1(t_9, t_{18} and t_{27}) under three different scenarios were all satisfied as shown in Figure 1S (supplement). However, in the case of update function for AKT, ACCORDION added a number of new AKT regulators to the baseline model which affected the dynamic behavior of AKT. Again, this demonstrates the dependance of ACCORDION output on the CE sets provided by machine reading. There are two ways in which this could be overcome. First, one could either use other tools to filter or score individual interactions in CE set [44][12] before they are used by ACCORDION, which we are planning to incorporate as one of our next steps. Second, ACCORDION can be used to identify cases where human input is necessary, for example, cases where many element regulators appear in literature, not all of which can be used to form regulatory rules.

3.3. Finding most relevant set of new interactions

We created the use cases such that the relationship between the number of elements and interactions in baseline models ($|V^{BM}|$, $|E^{BM}|$), and in their corresponding CE sets ($|V^{CE}|$, $|E^{CE}|$) varies, from the CE set being smaller than baseline model in the T-LGL Q^{Sm} case, to being up to six times larger than baseline model in other use cases (**Figure 2**). We also determined the size of the overlap, $|V^{BM} \cap V^{CE}|$ (see **Figure 2**), further highlighting that indeed the number of new elements that could be added to the model is much larger than the number of elements in the model.

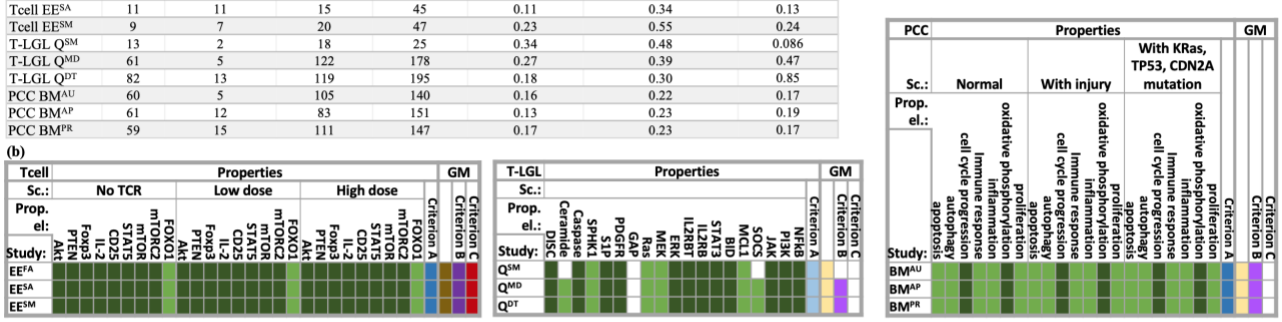
Additionally, we created these nine case studies such that they have baseline models with varying level of network connectivity. As described in Section 2S in the supplement, the baseline model in the T cell studies is a previously published, thus functional, model, while the T-LGL and PCC

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.22.477231; this version posted January 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(a)

STUDY	number of clusters	number of merged clusters	$\lvert E^{CM} \setminus E^{BM} \rvert_{avg}$	$\lvert E^{CM} \setminus E^{BM} \rvert_{max}$	$\left(\frac{\lvert E^{CM} \setminus E^{BM} \rvert}{\lvert E^{CE} \rvert}\right)_{avg}$	$\left(\frac{\lvert E^{CM} \setminus E^{BM} \rvert}{\lvert E^{CE} \rvert}\right)_{max}$	CM generating runtime [s]
Tcell EE^{FA}	22	5	26	81	0.10	0.33	0.09
Tcell EE^{SA}	11	11	15	45	0.11	0.34	0.13
Tcell EE^{SM}	9	7	20	47	0.23	0.55	0.24
T-LGL Q^{SM}	13	2	18	25	0.34	0.48	0.086
T-LGL Q^{MD}	61	5	122	178	0.27	0.39	0.47
T-LGL Q^{DT}	82	13	119	195	0.18	0.30	0.85
PCC BM^{AU}	60	5	105	140	0.16	0.22	0.17
PCC BM^{AP}	61	12	83	151	0.13	0.23	0.19
PCC BM^{PR}	59	15	111	147	0.17	0.23	0.17

(b)



GRAPH CONTENT PREREQUISITE FOR SATISFYING A PROPERTY:		
$v^{property} \in V^{BM}$ ($v^{property} \in V^{CE}$ or $v^{property} \notin V^{CE}$)	$v^{property} \notin V^{BM}$ and $v^{property} \in V^{CE}$	$v^{property} \notin (V^{BM} \cup V^{CE})$
***CRITERION A*: FIND ALL PROPERTY NODES IN V^{CE}?**		
$\forall v^{property}$: $v^{property} \in CM^{recommended}$	$\forall v^{property}, \exists CM_j$: $v^{property} \in V^{CM_j}$	$\exists v^{property}, \forall CM_j$: $v^{property} \notin V^{CM_j}$
GRAPH CONTENT PREREQUISITE FOR FINDING MISSING EDGES THAT ARE IN E^{GM} BUT NOT IN E^{BM}:		
$\forall e_i \in E^{GM} \setminus E^{BM}$: $e_i \in E^{CE}$	$(\exists\ but\ not\ \forall) e_i \in E^{GM} \setminus E^{BM}$: $e_i \in E^{CE}$	$\forall e_i \in E^{GM} \setminus E^{BM}$: $e_i \notin E^{CE}$
***CRITERION B*: CAN ACCORDION FIND ALL MISSING NODES?**		
$\forall v_i \in (V^{GM} \setminus V^{BM}) \cap V^{CE}$, $v_i \in CM^{recommended}$	$\forall v_i \in (V^{GM} \setminus V^{BM}) \cap V^{CE}$, $\exists CM_j$: $v_i \in V^{CM_j}$	$\exists v_i \in (V^{GM} \setminus V^{BM}) \cap V^{CE}$, $\forall CM_j$: $v_i \notin V^{CM_j}$
***CRITERION C*: CAN ACCORDION FIND ALL MISSING EDGES?**		
$\forall e_i \in (E^{GM} \setminus E^{BM}) \cap E^{CE}$, $e_i \in CM^{recommended}$	$\forall e_i \in (E^{GM} \setminus E^{BM}) \cap E^{CE}, \exists CM_j$: $e_i \in E^{CM_j}$	$\exists e_i \in (E^{GM} \setminus E^{BM}) \cap E^{CE}, \forall CM_j$: $e_i \notin E^{CM_j}$

Figure 4. (a) Characterization of CMs created by ACCORDION for the nine case studies. (b) Three criteria definitions and ACCORDION's criteria outcomes in the nine case studies ($v^{property}$ is the element included in the property as listed in Table 1S; $CM^{recommended}$ is the top recommended model).

baseline models were created by removing nodes and interactions from a published model. Since by construction the clusters that ACCORDION generates are usually connected only to a part of the baseline model, we used the node overlap metric *NO*, defined in Section 2.5, to determine the relationship between the number of new nodes that are added to the baseline model and the part of the model those nodes are connected to. The *NO* numbers in Figure 1S in the supplement, together with the ratios $\frac{|E^{CM} \setminus E^{BM}|}{|E^{CE}|}$ listed in **Figure 4**, show that ACCORDION is very selective, and it only adds to the baseline model a subset of new interactions that are well connected with the model.

We further investigated the percentage of these interactions selected from the entire CE set that were included in the top recommended CM (**Figure 4**(a)). For the Tcell cases, ACCORDION recommended on average 14% of the interactions as candidates for model extension, whereas for T-LGL and PCC cases, ACCORDION identified on average 26% and 15% of such interactions, respectively. These numbers emphasize an important characteristic of ACCORDION: while it provides a comprehensive overview of literature, it significantly reduces the number of selected interactions, such that, if human input is still necessary, the number of interactions to manually review is significantly smaller than the original CE set.

Interestingly, when observed together with the model error results, in the T cell and T-LGL studies, the higher *NO* values seem to correlate well with larger reduction in model error. However, in the PCC studies this correlation does not hold, where the CMs with a large number of new interactions compared to the size of the baseline model significantly decrease the baseline model error (~80% reduction). This demonstrates another important characteristic of ACCORDION: when the baseline model is already well-built, a smaller number of extensions can help improve it (e.g., Tcell and T-LGL cases), while for baseline models that are not very well connected and not functional or usable to start with (e.g., when the user starts only with a seed set of interactions and not a complete model), a larger number of interactions needs to be added to improve them (e.g., PCC case).

3.4. Identifying alternative networks

As described in Section 3.1, and also detailed in Section 2S in the supplement, we identified golden models for our case studies. Our goal with using golden models was twofold: we were interested in exploring how closely ACCORDION can reproduce previously published models ("criterion B" and "criterion C" in Figure 4) as well as comparing and contrasting them to automatically created models that satisfy the same set of properties.

In all three T cell case studies, ACCORDION adds all the interactions from the $E^{GM} \setminus E^{BM}$ set to its top recommended CMs (columns "GM" in Figure 4, dark yellow cells). For example, the merged cluster $C_{1,2}^{SM}$, with *NO*=0.7, restored all the missing interactions that were removed from the golden model. In the T-LGL and PCC studies, ACCORDION adds 30% and 32% of missing golden model interactions to recommended CMs. However, while in all three T cell studies all missing golden model interactions, i.e., interactions from the $E^{GM} \setminus E^{BM}$ set are present in CE sets, the CE sets in the T-LGL and PCC studies do not contain all the interactions from the $E^{GM} \setminus E^{BM}$ sets, as shown in Figure 4 (columns "GM", dark yellow cells). This is due to either papers that were selected using queries do not include those missing interactions or machine reading does not recognize these interactions in the papers.

An important outcome from this exercise is that ACCORDION recommends new CMs, different from golden models, which have high σ-score and δ-score and contain new interactions that form return paths with the baseline model. Moreover, in the T-LGL studies, a significant portion of interactions (41%) was removed from the golden model to obtain the baseline model. In such cases, ACCORDION selected from the large CE sets many additional interactions that form stronger connections with the baseline model (as part of clusters with high *NO* values and return paths) than the ones that are in the golden model, while also being able to find CMs that have high σ-score and δ-score. For instance, the regulators of AKT in the golden model are PIP3 and mTORC2, while the models recommended by ACCORDION also include regulations by TGFB, IFNgamma, CK2, CTLA4, SHIP1, all of which are suggested in literature.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.22.477231; this version posted January 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

This highlights another possible use of ACCORDION, when examining redundancies in signaling networks or discovering alternative pathways regulating the same target element.

3.5. Assistance in query answering

We also explored the relationship between the design of queries and ACCORDION's effectiveness, that is, whether the selection of search terms to mine literature affects the usefulness of extensions selected by ACCORDION. As described in the supplement, for the Tcell CE^{FA} case, we used a search query as an input to PubMed to identify the most relevant papers. We investigated the influence of this query on the percentage of interaction in clusters used to create CMs with top scores. In Figure 4, we show the average and the maximum percentage of selected interactions, i.e., $\left(\frac{|E^{CM}\setminus E^{BM}|}{|E^{CE}|}\right)_{avg}$ and $\left(\frac{|E^{CM}\setminus E^{BM}|}{|E^{CE}|}\right)_{max}$, which are 10% and 33%, respectively. For the best recommended model of this particular case study, ACCORDION was able to recover all the missing elements that are in V^{GM} and not in V^{BM}, namely, FOXO1, NEDD4, CK2 and MEK1. Furthermore, as can be seen in Figure 1S (supplement), ACCORDION recapitulated the dynamic behavior of FOXO1, an element that was in the search query used to collect interactions for the CE set (Section 2S), in all three scenarios (properties t_9, $t_{18}$and t_{27}). However, the dynamic behavior of AKT (also in the search query), IL2 and STAT5 was not recovered in one out of three scenarios, (high TCR scenario, properties t_{19}, $t_{22}$and t_{24}). This is due to potentially erroneous interactions in the CE set extracted by machine readers, e.g., CD8 → AKT, proliferation → AKT, differentiation -| AKT, differentiation -| IL2 and differentiation -| STAT5 ("→" represents positive regulation, "-|" represents negative regulation, also used in Figure 1). As mentioned above, we plan to add pre-processing of CE sets (e.g., using interaction filtering [44]).

For the T-LGL model study, we used three different queries as described in Section 2S in the supplement. The most elaborate query, in the T-LGL Q^{Det} case study, introduced more descriptive search terms, led to selecting more relevant papers, and consequently, extraction of relevant events and element regulators resulting in recommendation of a CM with high σ-score (0.76) and δ-score (0.75). Additionally, the update rules of most of the elements were retrieved except three elements, S1P, GAP and IL2RB. The properties that correspond to these three elements are properties t_5, t_7 and t_{12}. In contrast, for T-LGL Q^{Sm} and T-LGL Q^{Med} cases, less properties have been satisfied. For example, the baseline model error in property t_{17}, related to the behavior of element JAK, is not corrected in the T-LGL Q^{Sm} case, while property t_{19}, related to element NFκB, is not corrected in both T-LGL Q^{Sm} and T-LGL Q^{Med} cases. This is mainly due to the key regulatory interactions for these elements not being extracted from literature, or due to the interactions that are recovered not forming proper update functions. Overall, by comparing the results for the three queries in the T-LGL case studies, we have confirmed that a better query design leads to more useful and relevant information in the input CE sets.

3.6. Runtime

In Figure 4(a), we list the time that ACCORDION takes to generate clusters when run on a 3.3 GHz Intel Core i5 processor. The time required by ACCORDION to generate clusters increases with larger CE sets. For the PCC case studies, the runtime same across studies since the same CE set has been used. However, for the T cell and T-LGL case studies, the CE sets have different sizes, and thus, result in different runtime. The runtime of the overall extension algorithm is proportional to the number of properties that we need to test against. In other words, if we have N_C clusters and N_P properties, the time required for the extension algorithm is at the order of $O(N_C{*}N_P)$. However, the runtime can be significantly reduced if testing for all properties and clusters is carried out in parallel, which is part of our immediate future work.

4. Conclusions

In this paper, we have described a novel methodology and a tool, ACCORDION, that can be used to automatically assemble the information extracted from literature into models and to recommend models that achieve desired dynamic behavior. Our proposed approach combines machine reading with clustering, simulation, and model checking, into an automated framework for rapid model assembly and testing to address biological questions. Furthermore, by automatically extending models with the information published in literature, our methodology allows for efficient collection of the existing information in a consistent and comprehensive way, while also facilitating information reuse and data reproducibility, and often helping replace tedious trial-and-error manual experimentation, thereby increasing the pace of knowledge advancement. The results we presented here demonstrate different research scenarios where ACCORDION can be used. Both the benchmark set we presented, and the ACCORDION tool with detailed documentation are prepared for open access. As our next steps, we are planning to improve the input pre-processing in order to provide more useful candidate event sets, to make ACCORDION compatible with other model representation formats (e.g., SBML), as well as to work on parallelizing the tool implementation to improve the runtime when testing large number of properties.

Funding

This work was supported in part by DARPA grant W911NF-17-1-0135 awarded to N. Miskov-Zivanov.

References

[1] J. Epstein, "Why Model?," *Cybern. Syst.*, vol. 35, no. 2–3, pp. 117–128, 2008.

[2] A. Fabregat *et al.*, "The Reactome Pathway Knowledgebase," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D649–D655, 2018.

[3] C. Von Mering *et al.*, "STRING : known and predicted protein – protein associations , integrated and transferred across organisms," vol. 33, pp. 433–437, 2005.

[4] K. F. Aoki and M. Kanehisa, "Using the KEGG Database Resource," *Curr. Protoc. Bioinforma.*, vol. 11, no. 1, pp. 1–54, 2005.

[5] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan, "Causal protein-signaling networks derived from multiparameter single-cell data," *Science (80-.).*, vol. 308, no. 5721, pp. 523–529, 2005.

[6] J. Fisher, "Executable Biology Tackling the problem of complexity," 2012.

[7] J. Fisher and T. A. Henzinger, "Executable cell biology," *Nat. Biotechnol.*, vol. 25, no. 11, pp. 1239–1249, 2007.

[8] M. A. Valenzuela-Escárcega, G. Hahn-Powell, T. Hicks, and M. Surdeanu, "A Domain-independent Rule-based Framework for Event Extraction," *Proc. ACL-IJCNLP 2015 Syst. Demonstr.*, pp. 127–132, 2015.

[9] M. A. Valenzuela-escárcega *et al.*, "Large-scale Automated Reading of Scientific Cancer Literature Discovers New Cancer Driving Mechanisms," pp. 3–5, 2017.

[10] J. F. Allen and C. M. Teng, "Broad coverage, domain-generic deep semantic parsing," *AAAI Spring Symp. - Tech. Rep.*, vol. SS-17-01-, pp. 108–115, 2017.

[11] "Eidos: https://github.com/clulab/eidos."

[12] B. M. Gyori, J. A. Bachman, K. Subramanian, J. L. Muhlich, L. Galescu, and P. K. Sorger, "From word models to executable models of signaling networks using automated assembly," *Mol. Syst. Biol.*, vol. 13, no. 11, p. 954, 2017.

[13] L. Licata *et al.*, "SIGNOR 2.0, the SIGnaling Network Open Resource 2.0: 2019 update," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D504–D510, 2020.

[14] Y. Ahmed and N. Miskov-Zivanov, "Guided assembly of cellular network models from knowledge in literature," *2021 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, pp. 4458–4464, 2021.

[15] Y. Ahmed, A. A. Butchy, K. Sayed, C. Telmer, and N. Miskov-Zivanov, "New advances in the automation of context-aware information selection and guided model assembly," arXiv preprint

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.22.477231; this version posted January 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

arXiv:2110.10841, Accepted at the 13th International Workshop on Bio-Design Automation (IWBDA21)..

[16] K.-W. Liang, Q. Wang, C. Telmer, D. Ravichandran, P. Spirtes, and N. Miskov-Zivanov, "Methods to Expand Cell signaling Models using Automated Reading and Model Checking." Springer, Cham, pp. 145–159, 2017.

[17] K. Sayed, K. N. Bocan, and N. Miskov-zivanov, "Automated Extension of Cell Signaling Models with Genetic Algorithm," no. 1, pp. 5030–5033, 2018.

[18] Y. Ahmed, C. Telmer, and N. Miskov-Zivanov, "CLARINET: Efficient learning of dynamic network models from literature," *Bioinforma. Adv.*, 2021.

[19] W. F. Hawse *et al.*, " Cutting Edge: Differential Regulation of PTEN by TCR, Akt, and FoxO1 Controls CD4 + T Cell Fate Decisions ," *J. Immunol.*, vol. 194, no. 10, pp. 4615–4619, 2015.

[20] R. Zhang *et al.*, "Network model of survival signaling in T-cell large granular lymphocyte leukemia," *Proc. Natl. Acad. Sci.*, vol. 105, no. 42, pp. 16308–16313, 2008.

[21] C. A. Telmer *et al.*, "Computational modeling of cell signaling and mutations in pancreatic cancer," in *Proc. of Artificial Intelligence for Synthetic Biology (AI4SynBio), November 2019*, 2019.

[22] N. Miskov-Zivanov, M. S. Turner, L. P. Kane, P. A. Morel, and J. R. Faeder, "The duration of T cell stimulation is a critical determinant of cell fate and plasticity," *Sci. Signal.*, vol. 6, no. 300, pp. 1–16, 2013.

[23] F. Bianconi, E. Baldelli, V. Ludovini, L. Crinò, A. Flacco, and P. Valigi, "Computational model of EGFR and IGF1R pathways in lung cancer: A Systems Biology approach for Translational Oncology," *Biotechnol. Adv.*, vol. 30, no. 1, pp. 142–153, 2012.

[24] R. T. Pillich, J. Chen, V. Rynkov, D. Welker, and D. Pratt, *NDEx: A community resource for sharing and publishing of biological networks*, vol. 1558. 2017.

[25] R. J. Roberts, "PubMed Central : The GenBank of the published literature," vol. 98, no. 2, pp. 381–382, 2001.

[26] "https://www.ovid.com." .

[27] S. E. Schaeffer, "Graph clustering by flow simulation," *Comput. Sci. Rev.*, vol. 1, no. september 1969, pp. 27–64, 2007.

[28] S. Brohée and J. van Helden, "Evaluation of clustering algorithms for protein-protein interaction networks," *BMC Bioinformatics*, vol. 7, 2006.

[29] X. Lei, F. Wang, F. Wu, A. Zhang, and W. Pedrycz, "Protein complex identification through Markov clustering with firefly algorithm on dynamic protein – protein interaction networks," *Inf. Sci. (Ny).*, vol. 329, pp. 303–316, 2016.

[30] J. Vlasblom and S. J. Wodak, "Markov clustering versus affinity propagation for the partitioning of protein interaction graphs," *BMC Bioinformatics*, vol. 10, pp. 1–14, 2009.

[31] F. B. J and D. D, "Clustering by passing messages between data points," *Science (80-.).*, vol. 315, no. February, pp. 972–976, 2007.

[32] A. D. King, N. Pržulj, and I. Jurisica, "Protein complex prediction via cost-based clustering," *Bioinformatics*, vol. 20, no. 17, pp. 3013–3020, 2004.

[33] M. Blatt, S. Wiseman, and E. Domany, "Superparamagnetic clustering of data," *Phys. Rev. Lett.*, vol. 76, no. 18, pp. 3251–3254, 1996.

[34] G. D. Bader and C. W. V Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC Bioinformatics*, vol. 27, pp. 1–27, 2003.

[35] P. A. Gagniuc, *Markov Chains: From Theory to Implementation and Experimentation*. 2017.

[36] I. Mountasser, B. Ouhbi, and B. Frikh, "Parallel Markov-based clustering strategy for large-scale ontology partitioning," *IC3K 2017 - Proc. 9th Int. Jt. Conf. Knowl. Discov. Knowl. Eng. Knowl. Manag.*, vol. 2, no. Keod, pp. 195–202, 2017.

[37] K. Sayed, Y. H. Kuo, A. Kulkarni, and N. Miskov-Zivanov, "DiSH simulator: Capturing dynamics of cellular signaling with heterogeneous knowledge," *Proc. - Winter Simul. Conf.*, pp. 896–907, 2018.

[38] S. Andjelkovic and N. Miskov-Zivanov, "DiSH-trend: Intervention Modeling Simulator That Accounts for Trend Influences.," in *arXiv preprint arXiv:2107.01302*, 2021.

[39] Q. Wang, N. Miskov-Zivanov, B. Liu, J. R. Faeder, M. Lotze, and E. M. Clarke, "Formal Modeling and Analysis of Pancreatic Cancer Microenvironment," pp. 289–305, 2016.

[40] S. Kumar-Jha, E. M. Clarke, C. J. Langmead, A. Legay, A. Platzer, and P. Zuliani, "A Bayesian Approach to Model Checking Biological Systems," *Cmsb*, no. 2005, pp. 218–234, 2009.

[41] I. Tkachev and A. Abate, "Formula-free Finite Abstractions for Linear Temporal," *Proc. 16th Int. Conf. Hybrid Syst. Comput. Control*, pp. 283–292, 2013.

[42] N. Miskov-Zivanov, P. Zuliani, E. M. Clarke, and J. R. Faeder, "Studies of biological networks with statistical model checking: Application to immune system cells," *2013 ACM Conf. Bioinformatics, Comput. Biol. Biomed. Informatics, ACM-BCB 2013*, pp. 728–729, 2013.

[43] "http://www.biodesignlab.pitt.edu/accordion."

[44] E. Holtzapple, C. A. Telmer, and N. Miskov-Zivanov, "FLUTE: Fast and reliable knowledge retrieval from biomedical literature," *Database (Oxford).*, vol. 2020, pp. 1–17, 2020.

[45] K. Sayed, C. A. Telmer, A. A. Butchy, and N. Miskov-Zivanov, "Recipes for translating big data machine reading to executable cellular signaling models," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10710 LNCS, no. 1, pp. 1–15, 2018.