

## SUPPLEMENTARY MATERIAL

The supplementary material for the paper entitled “Automated Hypotheses Generation via Evolutionary Abduction” includes *textual material* and *artifacts*. Textual material is in the following Appendixes [A-C](#). Artifacts includes the source code (and the executable `.jar`) of the proposed algorithm and of all the implemented baselines, the experimental code, the datasets used for the experimentation, and the results reported in the main text and in the following appendixes. These are available at: <http://github.com/eva-iclr-2021/EVA>.

The following textual supplementary material is organized as follows. First, the algorithms of the three abduction operators are described (Appendix [A](#)). Appendix [B](#) describes the baseline strategies we have implemented to solve the causal problem by causal structure discovery algorithms followed by sampling. Appendix [C](#) reports the results of the tuning of the parameters used in the experimentation. These refer to both the EVA hyperparameters and to the size of the population used in the experimental study. A best and worst case for EVA are derived, then used in the final experimentation reported in the main text. Appendix [D](#) reports the results achieved by the three abductive operators of EVA, which together contribute to the overall performance of EVA. Appendix [E](#) reports the distribution of the distances of the last generation’s solutions, namely of the final solutions. In particular, the distributions of the average and of the best distance (over the distances of the final population’s solutions) are shown, as well as for the *relative distance*. Appendix [F](#) shows the best solution (namely, the solution with the best distance) of the populations at every generation, averaged over the 10 repetitions. Finally, Appendix [G](#) details the ASRS dataset, which, unlike the other datasets, is prepared from scratch starting from the ASRS database.

## A THE EVOLUTIONARY ABDUCTION OPERATORS

Algorithms [2-4](#) are the *factual*, *analogical* and *hypothetical cause* operators of EVA described in Section [3](#).

Algorithm [2](#) is the *factual* operator. It takes, as input, a solution  $\mathbf{x}$ , chosen by the selection operator (`select_factual`), all the different sources and targets (i.e., causes and effects) that are in the current population ( $S$  and  $T$ ), and considers the  $KB$  to build a new solution. To build the new solution  $\mathbf{x}'$ , first, a target  $t$  is selected from the list of all the targets in the current population (line 1). Selection of the target is done taking two targets randomly, measuring their “support” (number of occurrences in  $KB$ ) and taking the one with greater support or choosing randomly (with probability 0.5) one of the two if they have equal support. In essence, it is a binary tournament applied to single elements rather than to the whole solution. As for the sources, the same sources of  $\mathbf{x}$  are used (line 2). The operator applies three types of modifications: `add`, `modify` or `delete` actions. It considers two parameters to regulate the extent of changes and the desired novelty: an integer called *factual change index*  $\gamma_F > 0$  and a double called *factual novelty index*,  $\eta_F \in [0; 1]$ . The number of changes  $c$  to apply are selected randomly, with  $c \in [1; \gamma_F]$  (line 3). The type of change (`add`, `modify` or `delete`) is also selected randomly with equal chance for the three actions (line 5). In case of *add* or *modify* (which is a replacement of a source), the new source is selected from the set  $S$  with probability  $\eta_F$  or from the  $KB$  with probability  $1 - \eta_F$ . Selection of the sources to `add`, `replace` or `remove` (lines 7, 11, 14, respectively) is done by a variable-level binary tournament like the above-mentioned target selection, so as to favour the sources/targets contributing more to plausibility.

Algorithm [3](#) presents the *analogical* operator. To build the new solution  $\mathbf{x}'$ , the operator first selects a target from  $T$ , just like the *factual* operator algorithm (i.e., via a variable-level binary tournament). Then, it builds the set of sources, coupled with the chosen target, by extracting and reproducing *structural* features of the sources in  $\mathbf{x}$  (`extractConstraint`, line 2). Three source-level constraints are defined currently in EVA, which require the new solution  $\mathbf{x}'$  to have progressively stronger similarities with  $\mathbf{x}$ :

- *Cardinality constraint*: the number of sources of  $\mathbf{x}'$  is required to be the same as  $\mathbf{x}$ . The cardinality is a “proxy” indicator for the complexity of a solution, since more sources means co-occurrences of more causes together for an effect. The selection of the sources for  $\mathbf{x}'$  is done by considering an analogical novelty index:  $\eta_A \in [0; 1]$ . The source to be added is selected from the set  $S$  with probability  $\eta_A$ , or from the ontology  $\Omega$  with probability  $1 - \eta_A$ .

**Algorithm 2: factual\_operator( $x, S, T$ )**


---

**Input** :  $\mathbf{x}$ , the selected solution;  $S/T$ , all different sources/targets in the current population;  $\eta_F$ , Factual novelty index;  $\gamma_F$ , Factual change index

```

1  $t \leftarrow \text{selectTarget}(T)$ ;
2  $\mathbf{x}' = \{\mathbf{x}, t\}$ ;  $\triangleright$  initialize  $\mathbf{x}'$  with the same sources as  $\mathbf{x}$ , and target  $t$ 
3  $c \leftarrow \text{Rand}(1, \eta_F)$ ;  $\triangleright$  Number of changes
4 for  $i=1$  to  $c$  do
5    $a \leftarrow \text{Rand}(\text{add}, \text{modify}, \text{delete})$   $\triangleright$  Action to apply
6   if  $a=\text{add}$  then
7      $s \leftarrow \text{selectSource}(\eta_F)$ ;  $\triangleright \eta_F$ : Prob. to select from  $S$  or from  $M$ 
8      $\text{addSource}(\mathbf{x}', s)$ ;
9   if  $a=\text{modify}$  then
10     $\text{removeSource}(\mathbf{x}')$ ;
11     $s \leftarrow \text{selectSource}(\eta_F)$ ;
12     $\text{addSource}(\mathbf{x}', s)$ ;  $\triangleright$  with  $s$  different from removed source
13   if  $a=\text{delete}$  then
14      $\text{removeSource}(\mathbf{x}')$ ;
15 return  $\mathbf{x}'$ ;
```

---

**Algorithm 3: analogical\_operator( $x, P, S, T$ )**


---

**Input** :  $\mathbf{x}$ , the selected solution (from  $M_E$ ),  $P$ , population;  $S/T$ , all different sources/targets in the current population;  $\eta_A$ , Analogical novelty index

```

1 ;  $t \leftarrow \text{selectTarget}(T)$ ;
2  $[p, v_g, \sigma_{M_g}] = \text{extractConstraints}()$ ;  $\triangleright$  Extract #sources ( $p$ ), #sources per group ( $v_g$ ),  $\sigma_{M_g}$  per group
3 for  $i=1$  to  $p$  do
4    $s \leftarrow \text{selectSource}(\eta_A)$ ;  $\triangleright \eta_A$ : Prob. to select from  $S$  or from  $\Omega$ 
5    $\text{addSource}(\mathbf{x}', s)$ ;
6 while ( $v_g$  and  $\sigma_{M_g}$  constraints are not satisfied) do
7    $\text{replaceSource}(\mathbf{x}')$ ;  $\triangleright$  Adjust the solution to meet constraints
8 return  $\mathbf{x}'$ ;
```

---

**Algorithm 4: hypothetical\_operator( $x, S, T$ )**


---

**Input** :  $\mathbf{x}$ , the selected solution;  $S/T$ , all different sources/targets in the current population;  $\eta_H$ , Hypothetical cause novelty index;  $\gamma_H$ , Hypothetical cause change index

```

1  $t \leftarrow \text{selectTarget}(T)$ ;
2  $\mathbf{x}' = \{\mathbf{x}, t\}$ ;  $\triangleright$  initialize  $\mathbf{x}'$  with the same sources as  $\mathbf{x}$ , and target  $t$ 
3  $c \leftarrow \text{Rand}(1, \eta_H)$ ;  $\triangleright$  Number of changes
4 for  $i=1$  to  $c$  do
5    $a \leftarrow \text{Rand}(\text{add}, \text{modify}, \text{delete})$   $\triangleright$  Action to apply
6   if  $a=\text{add}$  then
7      $s \leftarrow \text{selectSource}(\eta_H)$ ;  $\triangleright \eta_H$ : Prob. to select from  $S$  or from  $\Omega$ 
8      $\text{addSource}(\mathbf{x}', s)$ ;
9   if  $a=\text{modify}$  then
10     $\text{removeSource}(\mathbf{x}')$ ;
11     $s \leftarrow \text{selectSource}(\eta_H)$ ;
12     $\text{addSource}(\mathbf{x}', s)$ ;  $\triangleright$  with  $s$  different from removed source
13   if  $a=\text{delete}$  then
14      $\text{removeSource}(\mathbf{x}')$ ;
15 return  $\mathbf{x}'$ ;
```

---

(line 4). Selecting from the ontology rather than from the current population means that the agent intends to exploit new concepts not previously observed, taking from the entire knowledge about the domain of interest.

- *Group membership constraint*: in many causal problems, the joint causes available to explain an effect can be grouped in homogenous subsets (e.g.: all stress-related causes in medical diagnosis to explain a disease; or, in hazard analysis, all the environment-related events possibly causing an accident). In such cases, this constraint requires the new solution  $\mathbf{x}'$  to have the same structure as  $\mathbf{x}$ , namely the same number of subsets with the same cardinalities. For each group in  $\mathbf{x}'$  there must be one distinct group in  $\mathbf{x}$  with the same cardinality and viceversa.
- *Ordinal constraint*, uses the notion of support referred to subsets of  $\mathbf{s}$  (rather than to entire solutions as in Def. 4):

**Definition 7 (Support of order  $k$  and maximum support).** The support of order  $k$ ,  $\sigma_k(\mathbf{q})$ , of a subset of sources  $\mathbf{q} \subseteq \mathbf{s}$ , with  $k \leq |\mathbf{q}|$ , is the number of distinct  $k$ -tuples of  $\mathbf{q}$  that occur at least once in the solutions of the memory  $M$ . The *maximum degree of support*,  $\sigma_M(\mathbf{q})$ , of  $\mathbf{q} \subseteq \mathbf{s}$  is the maximum value of  $k$  such that  $\sigma_k(\mathbf{q}) > 0$ .

The *ordinal constraint* requires  $\mathbf{x}'$  to have the same number of subsets with the same  $\sigma_M$  values of the subsets of  $\mathbf{x}$ . For each pair of groups  $g'_i, g'_j$  in  $\mathbf{x}'$  with  $\sigma_M(g'_i) > \sigma_M(g'_j)$ , there must be a distinct pair of group in  $\mathbf{x}$ ,  $g_i, g_j$  with the same relation,  $\sigma_M(g_i) > \sigma_M(g_j)$  and viceversa. For implementing constraint 2 and 3, the sources added to match constraint 1 (line 4 of the Algorithm) are replaced in those (pairs of) groups that violate constraint 2 and/or 3 (line 6-7), until the constraint is met or a maximum number of attempts is reached.

Algorithm 4 is the *hypothetical-cause* abduction operator. This operator mimics the creative abduction allowing a human to advance hypotheses exploiting just his knowledge about the domain of interest (i.e., the ontology  $\Omega$ ). The initial solution  $\mathbf{x}'$  is build like in the factual abduction. The operator also applies the same actions as the factual operator: *add*, *modify* or *delete*, again exploiting parameters to regulate the extent of changes and novelty (*hypothetical change index*  $\gamma_H$  and *factual novelty index*,  $\eta_H \in [0; 1]$ ). The main difference lies in considering  $\Omega$  in lieu of  $KB$  as set from which a source can be selected, thus opening to a wider range of novel solutions.

A consequence is that these solutions are expected to have higher novelty compared to the factual operator, contrasted by a lower plausibility. And this is what actually happens by adopting such two types of reasoning: while factual abduction supports more plausible but less original inference, hypothetical abduction, by its nature, is open to completely new scenarios but whose plausibility can be low. Analogical abduction lies in between; in fact, it is also called a *partially* ampliative inference. Although one can focus on just one of these operators in custom implementations of EVA, the suggestion is to exploit all the three operators for their complementarity.

## B GRAPH-BASED BASELINE STRATEGIES

The graph-based (GB) strategies have been implemented as follows. A Causal Structure Discovery (CSD) algorithm is used to learn the causal structure from the knowledge base  $KB$ ; the output is directed acyclic graph (DAG) with nodes being the variable and arcs being dependency relation between them (Pearl 2009). This is exploited to generate solutions proportional to cause-effect strength as described hereafter.

The CSD algorithms, namely FGES (Ramsey 2015), RFCI (Colombo et al. 2012), and GFCI (Ogarrío et al. 2016), are all present in the *py-causal* repository (Vowels et al. 2021) (PYCZEN), which exploits the Tetrad toolbox (Ramsey et al. 2018) tet. The parameters setting to derive the DAG and the corresponding arc weights are in Table 3 – the default parameters are kept, except the number of bootstraps (i.e., number of resampling) raised to 50 to improve the accuracy. The data type is always “discrete”. The description of each field can be found at <http://cmu-phil.github.io/tetrad/manual/>:

As prior knowledge, we specified (by the `priorKnowledge` parameter) that arcs between causes should be forbidden, as we are interested in arcs between causes and effects. The weights between arcs from causes to the effect obtained for the four datasets (values in the repository,

Table 3: ...

	FGES	GFCI	RFCI
scoreId	bdeu-score	bdeu-score	—
testId	—	disc-bic-test	bdeu-test
maxDegree/depth	3	3	3
faithfulnessAssumed	True	True	—
numberResampling	50	50	50
resamplingEnsemble	1	1	1
maxPathLength	—	-1	-1
completeRuleSetUsed	—	False	False
addOriginalDataset	True	True	True

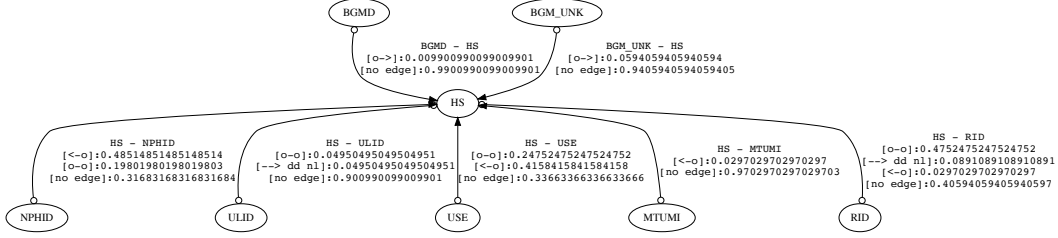


Figure 2: Example of DAG used for the RFCI GB strategy

<http://github.com/eva-iclr-2021/EVA>), which represent the probability that a potential *cause* node is causally related to the *effect* node, are used to generate the solution. An example of so-obtained DAG is in Figure 2 wherein *HS* is the effect (Hypoglycemic symptoms) and all the other variables are potential causes such as “More-than-usual meal ingestion” (MTUMI), “Blood Glucose Measurement Decrease” (BGMMD). This is obtained by the RFCI algorithm.

Given the graph and their cause-effect weights  $W = \{w_i\}$ ,  $i = 1, \dots, n$  and  $n$  being the number of (source) variables, the implemented generator acts as follows: for each instance to generate *i*) includes variable  $v_i$  ( $i = 1, \dots, n$ ) with probability  $w_i$  as part of the solution, and then *ii*) selects a value  $j$  of the variable  $v_i$ , say  $h_{i,j}$ , proportionally to the estimate of its probability of occurrence  $p_{i,j}$  within the *KB* – obtained as (normalized) relative frequency of that value within *KB*.

The random strategy just the variables  $v_i$  ( $i = 1, \dots, n$ ) with equal probability of selection, and then the values  $h_{i,j}$  of  $v_i$  with equal probability of selection.

## C PARAMETERS TUNING

A grid search approach is adopted for parameters tuning. The EVA hyperparameters are the novelty indexes,  $\eta_F$ ,  $\eta_A$  and  $\eta_H$ , and the change indexes  $\gamma_F$  and  $\gamma_H$  of the abduction operators. Both regulate the extent to which solutions are required to be diverse (hence novel) with respect to the *KB* and to the current population: the higher the  $\eta$ . values, the higher the probability of selecting new unseen sources, and the higher the  $\gamma$ . the higher the number of modifications that are done to build a (factual or hypothetical-cause) solution. The following configurations are considered:  $\langle \eta, \gamma \rangle = \langle 0.1, 3 \rangle, \langle 0.5, 5 \rangle, \langle 0.9, 7 \rangle$ , representing, respectively, a *Low* novelty degree in the solution, a *Medium* novelty and a *High* novelty.

Additionally, due to its evolutionary nature, EVA exploits the notion of population of solutions, whose *size* can impact the final results. Three values are considered for the population size:  $|P| = (15, 30, 60)$ .

We ran 10 repetitions for each of the  $3 \times 3 = 9$  configurations, each one for 600 evaluations, for the four datasets. Table 4 reports the average distance of the final population’s solution (averaged over the 10 repetitions) from the test set. The **best (B)** and **worst (W)** configurations for EVA are

Table 4: Average distance (standard deviation) of solutions of the best population – mean over 10 repetitions

		$ P  = 15$	$ P  = 30$	$ P  = 60$
<b>TUMOR</b>	Low	0.1680 <sub>0.0238</sub>	0.1821 <sub>0.0221</sub>	0.2118 <sub>0.0155</sub>
	Medium	0.1648 <sub>0.0240</sub>	0.1755 <sub>0.0154</sub>	0.2099 <sub>0.0205</sub>
	High	0.1620 <sub>0.0252</sub>	0.1652 <sub>0.0335</sub>	0.2034 <sub>0.0332</sub>
<b>ASRS</b>	Low	0.5407 <sub>0.0259</sub>	0.5833 <sub>0.0118</sub>	0.6376 <sub>0.0185</sub>
	Medium	0.4803 <sub>0.0203</sub>	0.5204 <sub>0.0181</sub>	0.5961 <sub>0.0191</sub>
	High	0.5162 <sub>0.0178</sub>	0.4971 <sub>0.0314</sub>	0.5560 <sub>0.0196</sub>
<b>MEDICAL</b>	Low	0.3584 <sub>0.0209</sub>	0.3651 <sub>0.0197</sub>	0.3631 <sub>0.0175</sub>
	Medium	0.3629 <sub>0.0212</sub>	0.3421 <sub>0.0169</sub>	0.3677 <sub>0.0134</sub>
	High	0.3557 <sub>0.0341</sub>	0.3582 <sub>0.0179</sub>	0.3615 <sub>0.0108</sub>
<b>NURSERY</b>	Low	0.0742 <sub>0.0479</sub>	0.0806 <sub>0.0306</sub>	0.1292 <sub>0.0193</sub>
	Medium	0.0850 <sub>0.0270</sub>	0.1101 <sub>0.0309</sub>	0.1517 <sub>0.0297</sub>
	High	0.1253 <sub>0.0187</sub>	0.1336 <sub>0.0300</sub>	0.1723 <sub>0.0196</sub>

highlighted (green and red, respectively). These two configurations are used to compare EVA with the baselines (over 6,000 evaluations) (cf. with Section 6), considering both the best and the worst case.

## D RESULTS BY EVA OPERATOR

Figure 3 reports the average distance of the final solutions computed by each of the three operators of EVA (i.e.: *Factual*, *Analogical*, *Hypothetical-cause*), in every run and experimental scenario (10 runs per scenario) for every dataset.

Two main observations arise: *i*) the *Factual* and *Hypothetical-cause* abduction operators give similar distance values for all scenarios and datasets. These, in fact, have the same structure, the main difference is in the source of knowledge used (the former relies on the *KB*, while the latter on the ontology  $\Omega$ ); *ii*) the *Analogical* operator works better (i.e., small distances) than the others for small problems, namely when few multiple causes are involved, which is the case of the MEDICAL and NURSERY datasets; in contrast *Factual* and *Hypothetical-cause* outperform the *Analogical* operator for ASRS and TUMOR. The impact of the Best/Worst configuration is negligible, as it does not change the relative results. A higher novelty constraint up to  $\nu_0 = 0.7$  causes the operators’ results to flatten on values above 0.6, as it becomes difficult for all the operators to find close-to-real solutions that are also very different from the *KB*. The only exception is the case of NURSERY, where the analogical operator still manage to give solutions with distance around 0.5 even with such a strict constraint on the novelty.

Although in one specific problem one operator may provide better solutions, for EVA to work reasonably well with various problems of different size, the suggested strategy is to always exploit the contribution of all the three operators. This also ensures a better diversity of the obtained solutions.

## E DISTRIBUTION OF SOLUTIONS

Figure 4 reports the percentage of solutions of the final generation’s population with average distance less than or equal to a given value – the average over 10 repetitions is reported. For the baseline strategies, since there is no notion of “evolution” and runs (i.e., generations) are independent of each other, we do not consider the final generation, but select the generation with the best population (i.e., having solutions with the best average distance). Results are broken down by novelty constraint and by configuration (best: **B**, worst: **W**).

EVA generates considerably more solutions in the left side of the histogram (i.e., closer to 0) for all the cases. In terms of datasets, the gain is more evident for more complex problems (ASRS, TUMOR), but also for problems with a may instances in the test set (NURSERY), while it becomes less evident for MEDICAL. Again, the Best/Worst configuration makes no relevant difference. With

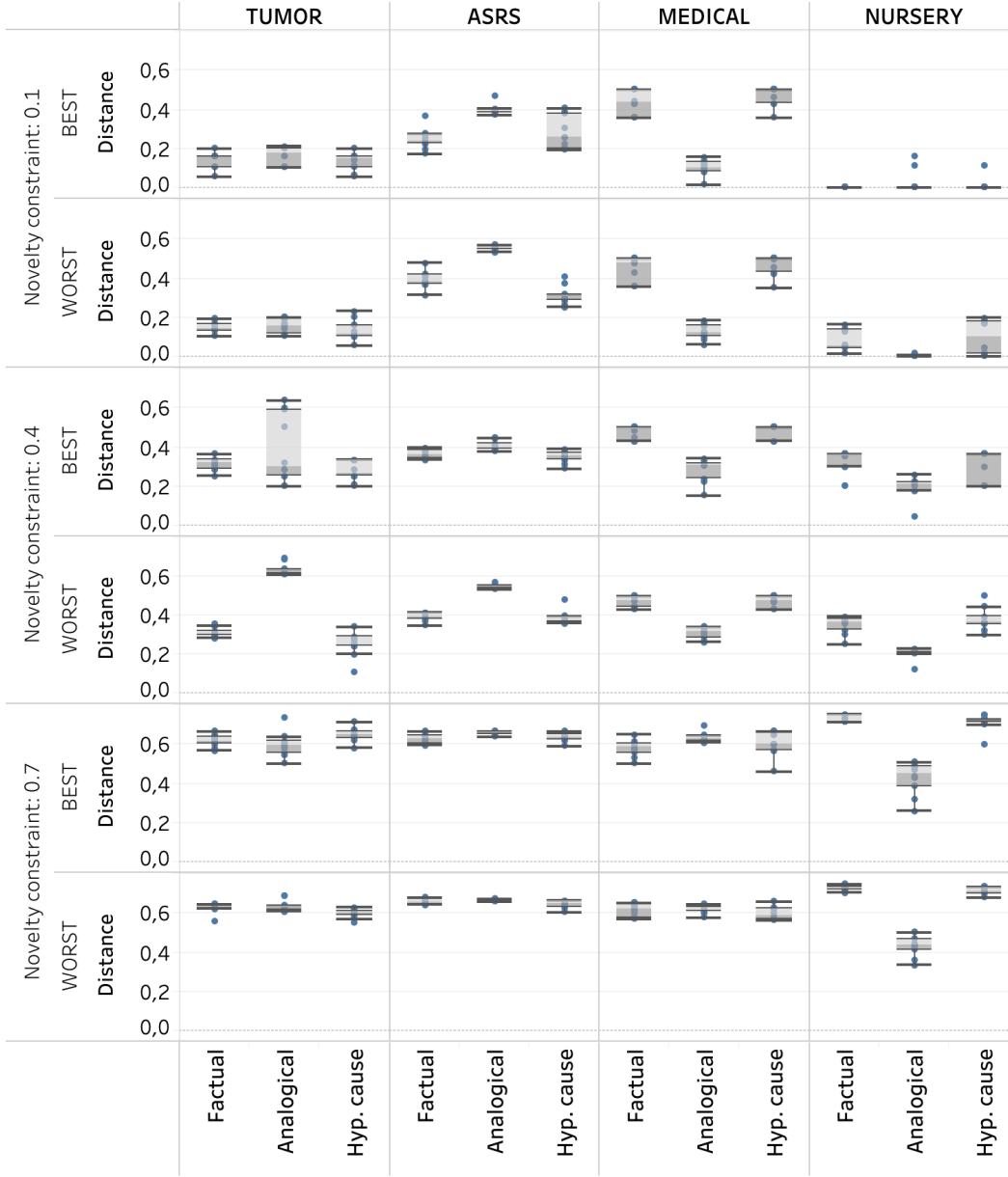
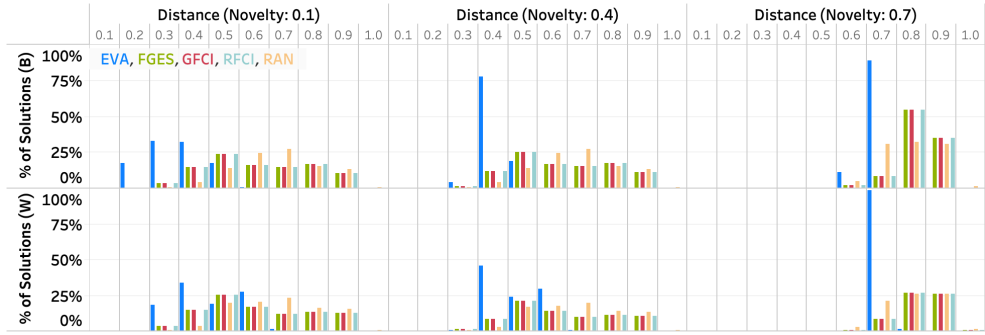


Figure 3: Results by operators

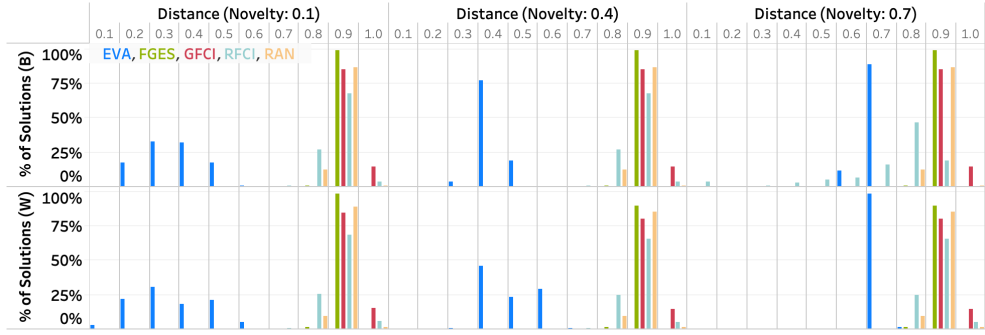
the increase of the novelty constraint the gain of course reduces, as there is less margin for improving over a random or graph-based strategy.

Figure 6 reports the same results but for the *relative* distance (cf. with Section 6). There are many cases in which solutions with relative distance equal to 0 are generated, namely solutions in which the set of causes is entirely contained in the set of causes of a real occurred event (an entry in the test set). For instance, in the MEDICAL dataset, many of the generated solutions (by all the techniques) have relative distance equal to 0<sup>5</sup>. For what said in Section 6, the gain of EVA in terms of relative distance is when the novelty constraint is at 0.1 and 0.4, not at 0.7.

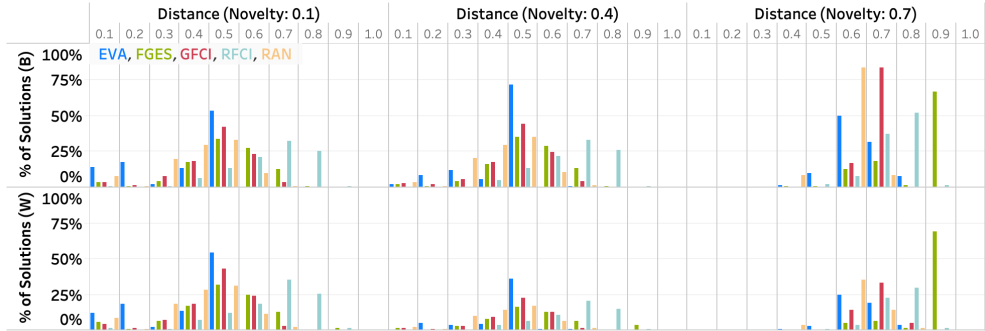
<sup>5</sup>Note that the objective is not to generate solutions with small relative distance, in which case would be enough to generate small solutions, e.g., with one single cause. The objective is to generate solutions with small absolute distance; this graph shows how often the so-generated solutions have small relative distance.



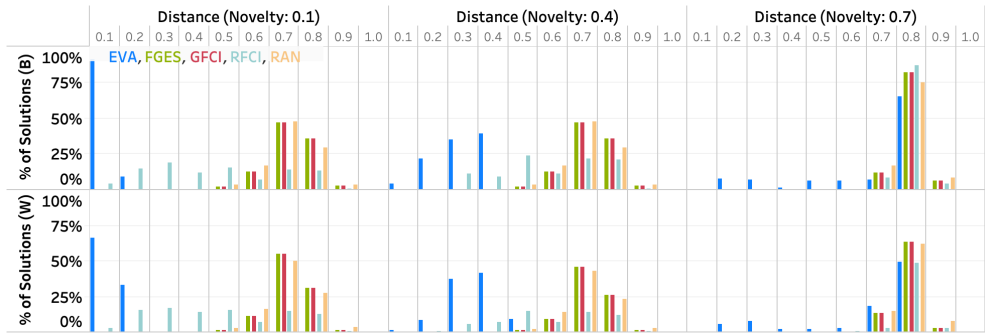
(a) TUMOR problem



(b) ASRS problem

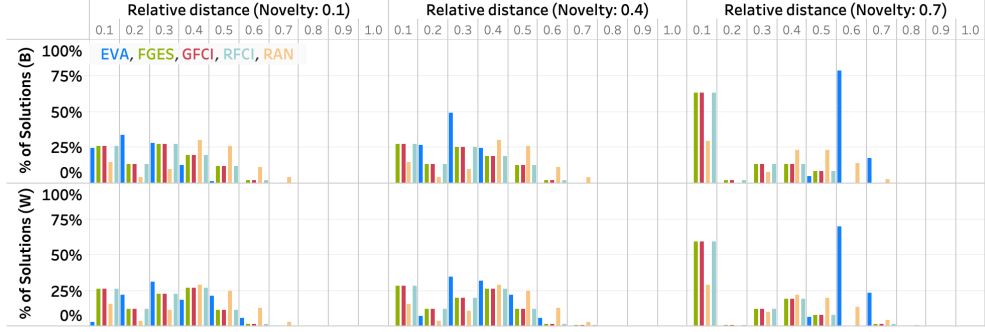


(c) MEDICAL problem

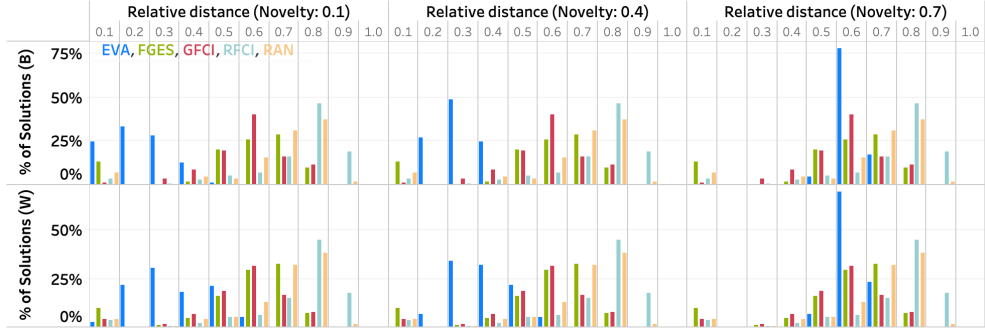


(d) NURSERY problem

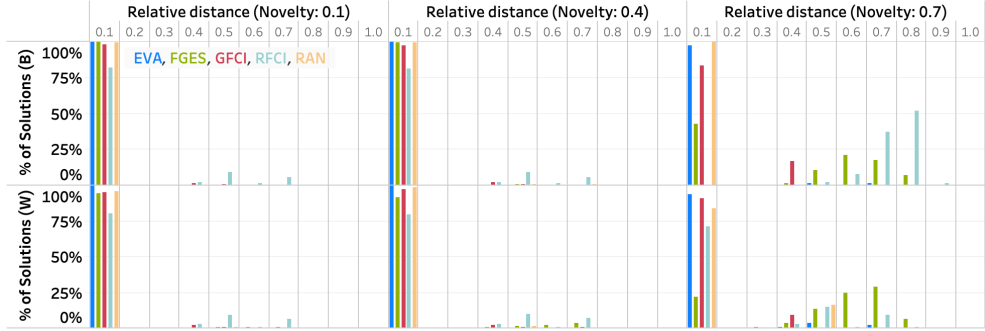
Figure 4: Distribution of solution's distance



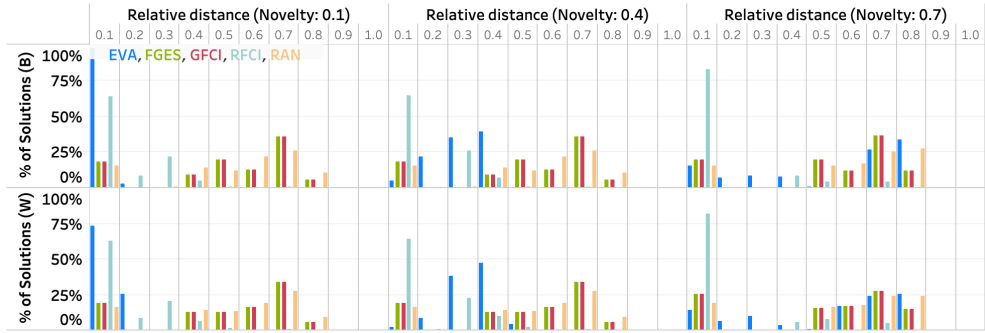
(a) TUMOR problem



(b) ASRS problem



(c) MEDICAL problem



(d) NURSERY problem

Figure 5: Distribution of solution's best distance

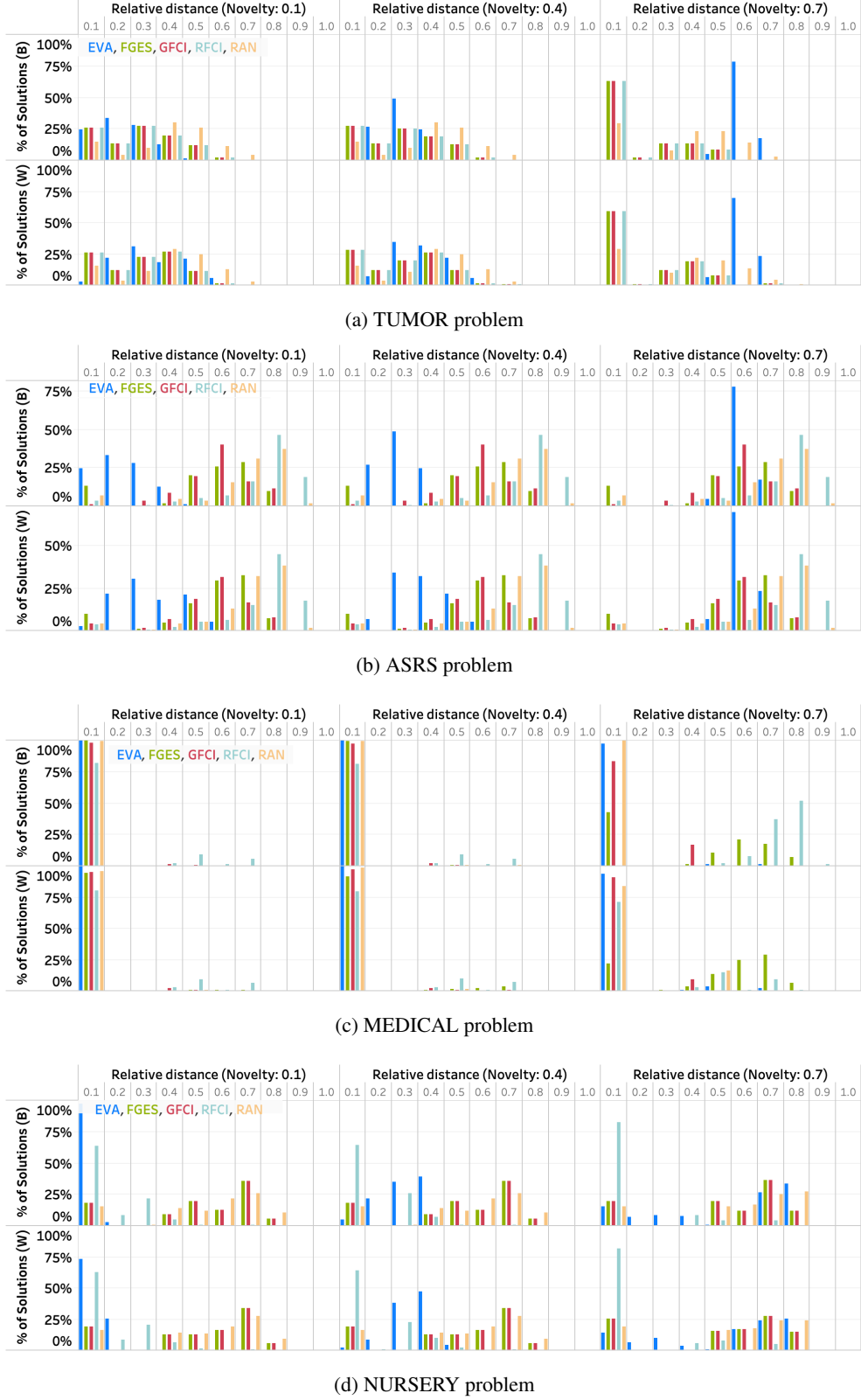
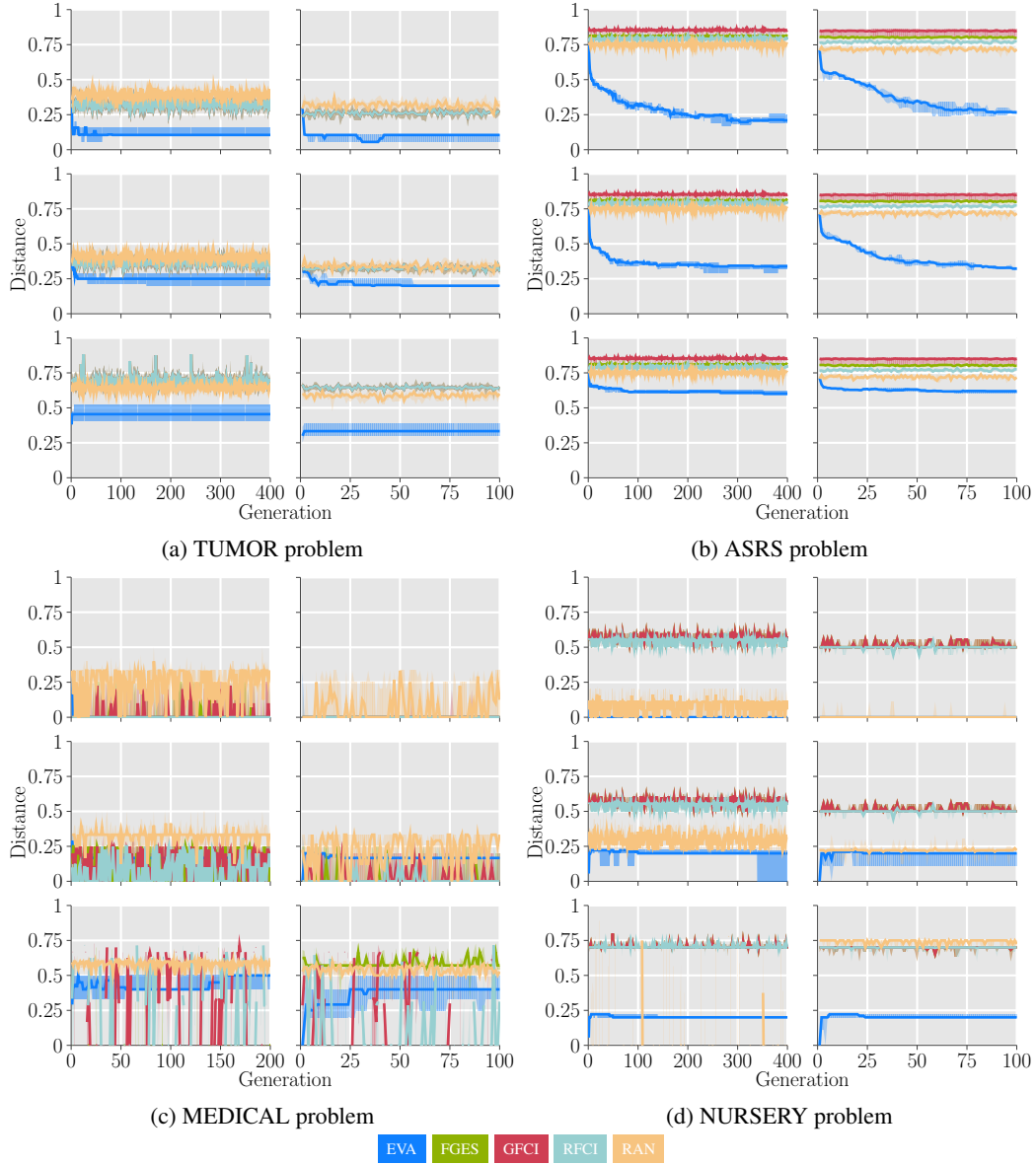


Figure 6: Distribution of solution's relative distance

Figure 7: *Best distance by generation.*

## F BEST SOLUTIONS BY GENERATION

Figure 7 reports the *best* distance of the population’s solutions vs. generations (median and IQR over 10). These are the same type of graph as Figure 1 in Section 6 but here the best solution of the population at every generation is considered. The evolution across generations leads to the final results summarized in Table 1 in Section 6. The distances are of course smaller than the average distances of Figure 1. In the case of ASRS, EVA gives distances that still decrease after 6,000 evaluations – it can still improve in that case, while in other cases it converged. When EVA is not visible in the graph (e.g., MEDICAL and NURSERY) it means the distances are 0. Finally, in the case of  $\nu_0 = 0.7$ , it often happens that the baselines do not provide solutions for some generations.

Table 5: ASRS. *Environment* entity.

Environment					
Flight conditions	Weather Elements/ Visibility	Work Env. Factors	Light	Ceiling	
VMC	Cloudy	Poor lighting	Dawn	CLR	
IMC	Fog	Glare	Daylight	Single value	
Mixed	Hail	Temperature extreme	Dusk		
Marginal	Haze-Smoke	Excessive humidity	Night		
	Icing				
	Rain				
	Snow				
	Thunderstorm				
	Turbulence				
	Windshear				
	Other				

Table 6: ASRS. *Aircraft* entity.

Aircraft						
Flight plan	Flight Phase	Route in use	Navigation in use	Cabin Lighting	Maintenance status & items	Mission
VFR	Taxi	Direct	FMS/FMC	High	Deferred	Aerobatics
IFR	Parked	Oceanic	GPS	Medium	Records	Agriculture
SVFR	Takeoff	VFR Route	INS	Low	complete	Ambulance
DVFR	Initial	Vectors	Localizer/	Off	Released	Banner tow
None	climb	Visual appr.	Gideslop/ILS		for serv.	Ferry
	Climb	None	NDB		Required	Cargo/Freight
	Cruise	Airway	VOR/VORTAC		Scheduled	Passenger
	Descent	STAR			Unscheduled	Photo shoot
	Initial Appr.	SID				Personal
	Final Appr.	Other			<i>Maintenance items</i>	Refueling
	Landing				Inspection	Skydiving
	Other				Installation	Tactical
					Repair	Test Flight
					Testing	Traffic watch
					Work cards	Training
						Utility
						Other

## G THE ASRS DATASET

While the TUMOR, MEDICAL and NURSERY datasets were already publicly available and explained, the ASRS dataset is new. Here we briefly describe the source of information from which the dataset is derived.

The Aviation Safety Reporting System (ASRS) database is the world’s largest repository of voluntary, confidential safety information provided by aviation personnel, including pilots, controllers, mechanics, flight attendants and dispatchers [ASR \(a\)](#).

It contains more than 1 million of entries reported since 1988. It is a structured database used for data retrieval and analysis, with all the accidents stored in a cause-effect style: the events regarding the aircraft components, the weather conditions, the human personnel involved, the airport, and many other potential causes recorded for each accident as a categorised set of values (i.e., enumerative), along with the resulting accident (also categorised). The main entities are reported in the following:

- **Environment**, with information regarding the flight conditions when accident occurred, visibility, working environment factors such as lighting or temperature.

Table 7: ASRS. *Component* entity

Component			
Component		Problem	
Indicating and Warning - Landing Gear	Weather Radar	Electrical Wiring & Connectors	Design
	DC Battery	Autopilot	Failed
	Turbine Engine	Landing Gear	Improperly operated
	Nose Gear	Yaw Control	Malfunctioning
	Flap Vane	Brake System	
	Powerplant Fire Extinguishing	Wheels/Tires/Brakes	
	Cockpit Window	Aircraft Cooling System	
	Turbine Assemb Blade	Landing Gear Indicating System	
	Normal Brake System	Tires	
	Gear Down Lock	Fuel System	
	Engine Control	Fire/Overheat Warning	
	Antiskid System	Piston	
	Fuselage Skin	Powerplant Fuel Control	
	External Power	Flap Control	
	Supplemental Landing Gear	FCC (Flight Control Computer)	
	Fuselage Panel	(more than 350)	
	Engine	...	

- **Aircraft**-related elements, e.g., the flight plan, the route, the flight phase, the maintenance status, the mission.
- **Component**, with information about all the components of the aircraft and their status (e.g., design problem, failed, malfunctioning).
- **Person**, reporting the information about the persons involved, such as the flight crew, the air traffic control, or people working in maintenance, information about the human factors that could cause mistakes such as distraction, confusion, stress, etc.
- **Events**, including anomalies such as airspace violation, deviation of altitude, procedural errors, airborne or ground conflict, fire, as well as the event describing the final result, such as the type of accident and its consequences (which correspond to our target variables).

An excerpt of the main information is reported in the Tables 5-9. A glossary of terms is available on the website ASR (b). For illustrative purpose, a solution looks like follows:

```
Environment.Weather = Fog
Environment.Weather = Windshear
Environment.Weather = Turbulence
Environment.FlightConditions = IMC
Environment.Light = Night
Aircraft.Mission = Cargo/Freight
FlightAircraft.Phase = Final Approach
Anomaly.Inflight Event = Object encountered
Result.Flight Crew = Landed in
Emergency Condition
Result.Aircraft = Aircraft Damaged
```

This describes an accident in which the pilot, while descending to approach for landing (Final Approach) during the night and under bad weather conditions (IMC stands for Instrument Meteorological Conditions as opposed to Visual Meteorological Conditions), struck a tree branch (Object encountered) and damaged the wing. Hence, he diverted to another airport, landing there in emergency conditions. This type combination is what EVA aims to construct by its operators as described in the main article. The dataset is made publicly available in our repository, <http://github.com/eva-iclr-2021/EVA>.

Table 8: ASRS. *Person* entity

<b>Person</b>			
Function	Qualification	Experience	Human Factors
<i>Flight crew</i>			
Captain	Student	Total	Communication breakdown
Check Pilot	Sport	Last 90 days	Confusion
First Officer	Private		Distraction
Flight Engineer	Commercial		Fatigue
Instructor	Air Transport Pilot		Human-Machine Interaction
Pilot Flying	Flight Instructor		Physiological
Pilot not Flying	Multiengine		Situational Awareness
Relief Pilot	Instrument		Time Pressure
Single Pilot	Flight Engineer		Training/Qualification
Trainee	Rotorcraft		Workload
Other	Lighter-Than-Air		Other
	Sea		
	Glider		
<i>Air Traffic Control</i>			
Approach	Fully certified	Radar	Location in aircraft
Coordinator	Developmental	Non-radar	Flight deck
Departure		Military	Cabin Jumpseat
Enroute		Supervisory	Crew Rest Area
Flight data			Doee Area
Flight service			Galley
Ground			General Searing Area
Handoff			Lavatory
Instructor			Other
Trainee			
Local			
Oceanic			
Supervisor			
Traffic Management			
Other			
<i>Maintenance</i>			
Inspector	Airframe	Avionics	
Instructor	Powerplant	Inspector	
Lead Technician	Appentice	Lead Technician	
Parts/Stores Personnel	Avionics	Repairman	
Quality Assurance	Inspection Authority	Technician	
Technician	Nondestructive Testing		
Trainee	Repairman		
Other			

Table 9: ASRS. *Events* entity

Events		
Anomalies	Assessment Primary or Contributory factor	Results
<i>Aircraft Equipment</i>		<i>General</i>
Critical	Aircraft	Declared Emergency
Less severe	Airport	Evacuated
	Airspace structure	Flight Cancelled/Delayed
<i>Airspace Violation</i>	ATC Equip	Maintenance Action
All types	/Nav Facility/Buildings	Physical Injury/Incapacitation
<i>ATC Issues</i>	Chart or Publication	Police/Security Involved
All types	Company Policy	Release Refused/Aircraft not Accepted
<i>Flight Deck/Cabin/Aircraft</i>	Equipment/Tooling	Work Refused
Illness	Env. non-weather related	None
Passenger Electronic Device	Human Factors	<i>Flight crew</i>
Passenger Misconduct	Incorrect/Not Instal.	Reoriented
Smoke/Fire/Fumes/Odor	/Unav. Part	Diverted
Other	Logbook Entry	FLC Overrode Automation
<i>Conflict</i>	Manuals	FLC Complied
NMAC	MEL	Executed Go Around/Missed Approach
Airbone conflict	Procedure	Exited Penetrated Airspace
Ground Conflict, critical	Staffing	Inflight Shutdown
Ground Conflict, less severe	Weather	Landed as Precaution
<i>Deviation - Altitude</i>		Overcame Equipment Problem
Crossing Restriction Not Met		Regained Aircraft Control
Excursion from Assigned Altitude		Rejected Takeoff
Overshoot		Requested ATC Assistance/Clarification
Undershoot		Returned to Clearance
<i>Deviation - Speed or Track/Heading</i>		Returned to Departure Airport
All types		Returned to Gate
<i>Deviation - Procedural</i>		Took Evasive Action
Clearance		<i>Air Traffic Control</i>
FAR		Provided Assistance
Hazardous Material Violation		Issued Advisory/Alert
Landing without Clearance		Issued New Clearance
Maintenance		Separated Traffic
MEL		<i>Aircraft</i>
Published Material/Policy 5205 - Security		Aircraft Damaged
Weight and Balance		Automation Overrode Flight Crew
Other/Unknown		Equipment Problem Dissipated
<i>Ground Excursion/Incursion</i>		
Ramp		
Runaway		
Taxiway		
<i>Ground Event/Encounter</i>		
Aircraft		
FOD		
Gear Up Landing		
Ground Strike Aircraft		
Loss of Aircraft Control		
Object		
Person/Animal/Bird		
Vehicle		
Other		
<i>Inflight Event/Encounter</i>		
CFTT/CFIT		
Fuel Issue		
Loss of Aircraft Control 5215 - Object		
Bird/Animal		
Unstabilized Approach		
VFR in IMC		
Wake Vortex Encounter		
Weather/Turbulence		