Evaluating Planning Model Learning Algorithms Across Different Representations

Anonymous Authors

Abstract

Automated planning is a prominent approach to sequential decision-making. A crucial aspect of domain-independent planning is the domain model, which provides to a planning engine the necessary application knowledge needed to synthesize solution plans. A domain model typically includes a state representation and an action model, which defines the set of possibly actions and the preconditions and effects of each action. Formulating domain-models is a challenging, time consuming, and error-prone task. For this reason, a number of approaches have been proposed to automatically learn complete (or partial) domain models from a set of provided observations. This raises the question of how to compare models learned by different approaches; there is a lack of evaluation metrics, and the complexity is exacerbated since learning algorithms may output models with different representations of states and actions. To foster the use of model learning approaches, in this paper we describe a set of metrics designed to assess different characteristics of models to be compared. Further, to bridge the potential representation gap between different learned models, we propose an encoder-decoder mechanism that allows to map two models regardless of their encoding, and we demonstrate how this encoder-decoder mechanism can be leveraged on to systematically compare models using the proposed metrics. Finally, suggest a benchmark suite based on existing domain models from the International Planning Competition (IPC) and an evaluation process for using it.

1 Introduction

Domain-independent planning is a foundational area of research in Artificial Intelligence (AI) that focuses on the automatic generation of plans to achieve specific goals from a given initial state in a given environment. Classical planning, which is the focus of this work, is the colloquial name for a well-studied type of domain-independent planning in which a single agent is acting in a fully observable, discrete, and deterministic environment. Most research on classical planning has focused on developing efficient algorithms for solving planning problems, and assumed the existence of a *domain model* specified in a formal language such as the Planning Domain Definition Language (PDDL) (Haslum et al., 2019). The domain model in classical planning defines how states are represented, the set of possible actions, and the preconditions and effects of each action. However, creating a domain model is a challenging, time-consuming, and error-prone task (McCluskey, Vaquero, and Vallati, 2017). This is a bottleneck for the wider dissemination of planning technology in real-world applications.

To address this issue, a number of algorithms have been proposed to automatically learn domain models from a set of provided observations (Gösgens, Jansen, and Geffner, 2025; Lamanna et al., 2025; Xi, Gould, and Thiébaux, 2024; Mordoch et al., 2024; Juba, Le, and Stern, 2021; Lamanna et al., 2021; Cresswell and Gregory, 2011; Zhuo et al., 2010).[Roni: Everyone is invited to add their action model learning algorithm here.] [Christian: Why not just point to some of the survey stuff? On the MACQ website (https://macq.planning.domains/), we have the bib entry and pointer to 3 previous survey papers. I'm biased, but I think MACQ should be included :P.] [Roni: Good idea, and I definitely think MACQ should be included. This is a mistake on my end, but will be remedied] Despite a recent resurgence in interest in learning domain models, there is no standard evaluation process for such algorithms, no set of agreed-upon evaluation metrics, and no standard benchmark. This paper aims to close this gap, and proposes an evaluation paradigm for domain model learning algorithms, a set of metrics, and a publicly available benchmark for evaluation.

We begin our exposition by describing a straightforward evaluation process for action model learning algorithms, which is based on comparing the *syntactic similarity* of the learned domain model to a *reference domain model*. We discuss the limitations of this evaluation method and propose an alternative evaluation process that aims to evaluate the *predictive power* and *problem-solving* ability of the learned domain model. Several specific evaluation metrics are defined for this type of evaluation based on prior works (Aineto, Celorrio, and Onaindia, 2019; Juba, Le, and Stern, 2021; Mordoch et al., 2024; Oswald et al., 2024), and we discuss their strengths and weaknesses.

The proposed metrics are designed for evaluating models that use the same representation of states and actions. However, in practice, action model learning algorithms may output models that use different representations of states and actions. For example, some algorithms output models in a grounded representation (Stern and Juba, 2017), while others output models in a lifted representation (Aineto, Celorrio, and Onaindia, 2019; Juba, Le, and Stern, 2021; Xi, Gould, and Thiébaux, 2024). Some represent states based

only the parameters of executed actions (Cresswell and Gregory, 2011), or different languages to support human engineers (McCluskey et al., 2010). Finally, some require a richer symbolic representation of states (Juba, Le, and Stern, 2021), while others use a visual representation of states (Xi, Gould, and Thiébaux, 2024). [Roni: All: add references and refine the above and add more examples of representation gap.] We propose an evaluation paradigm that addresses this representation gap challenge. In this paradigm, an evaluated domain model learning algorithm is obliged to define encoder and decoder functions to bridge the representation gap. We adapt the domain model evaluation metrics defined above to use these bridging functions, and show how to define such functions for several domain model learning algorithms. [Roni: TODO: Benchmarks, evaluation methodology] Finally, we describe a benchmark suite and suggested evaluation process for using it based on existing domain models from the International Planning Competition (IPC).

2 Background

We focus on *classical planning* problems, which is a well-studied type of planning problem in which a single agent is acting in a fully observable, discrete, and deterministic environment. A classical planning problem is defined as a tuple $\mathcal{P} = \langle F, A, s_0, G \rangle$, where F is a finite set of fluents, A is a finite set of actions, $s_0 \in S$ is the initial state, and $G \subset F$ is a set of fluents. [Roni: For all: I modified the above definition to:(1) not have S as the set of states but have F as the set of fluents, and (2) not have G as a set of states but as a subset of fluents (that must be achieved). If anyone objects please do not change but rather comment here on the change you would like and why.] A *state* is defined by a set of propositions, representing that the conjunction of fluents in this set are true in this state. An action $a \in A$ is defined as a tuple $a = \langle pre(a), \mathit{eff}(a) \rangle$, where $pre(a)$ is the precondition of a and $\mathit{eff}(a)$ is the effect of a. The precondition $pre(a)$ specifies the conditions that must hold in a state for the action a to be applicable, while the effect $\mathit{eff}(a)$ specifies the changes to the state resulting from applying the action a. The precondition and effect of an action are defined each as a set of literals, which are either positive or negative propositions. We denote by L as the set of literals, i.e., $L = \{\ell | \exists f \in F : \ell = f \vee \ell = \neg f\}$. [Roni: Not sure about the] A solution to a classical planning problem is a sequence of actions that transforms the initial state s_0 into a goal state s_g where $G \subseteq s_g$. [Roni: Minor modification to goal state to reflect the change above]

Planning domains are usually represented in a lifted representation, where the actions are defined with respect to a set of object types. The lifted representation of a planning domain is defined as a tuple $\mathcal{D} = \langle O, P, A \rangle$, where O is a set of object types, P is a set of predicates, and A is a set of actions. Actions and predicates are parameterized by objects, and preconditions and effects are defined accordingly. Popular classical planning systems, such FastDownward (Helmert, 2006), support such a lifted representation.

Different algorithms have been proposed for learning domains for classical planning. The input to these algorithms is a set of *trajectories*. A *trajectory* is a sequence of observations and actions. An *observation* can be a state or some other information about the state, such as a set of predicates that hold in the state or a visual representation of a state. Domain model learning algorithms differ in the type of trajectories they can learn from, the type of action models they can learn, and the representation of the learned action model.

[Christian: I'm actually not sure any of this is needed. We're probably pushing over 50 related works in the area (cf. MACQ or similar surveys), and there's no way we'll cover the full space of approaches. A few exemplary ones, and then a pointer to a survey should be fine. Unless we need the background of how some of these work, it's about half-a-page of space we could save. Now the *metrics* used in those papers, is something related.] [Roni: I agree this needs to be shortened. TODO for me. Would be very good to have a list of the metrics used in each paper. Great idea. TODO by someone? a table or so with metric and list of papers using it?]

The ARMS algorithm (Yang, Wu, and Jiang, 2007) requires as observations the initial and final state of each of the provided trajectories, with optional use of intermediate states if available. For each observed action, ARMS lifts the action and constructs a set of constraints on the fluents involved in its preconditions and effects. These constraints are then resolved using a weighted MAX-SAT solver to generate the action model with the highest weight. The Simultaneous Learning and Filtering (SLAF) (Amir and Chang, 2008) learns lifted action models from trajectories even if some of the states in them are not observed. SLAF uses logical inference to filter inconsistent action models and requires observing all actions in the trajectory. FAMA (Aineto, Celorrio, and Onaindia, 2019) can also handle missing observations and outputs a lifted planning domain model. It frames the task of learning an action model as a planning problem, ensuring that the returned action model is consistent with the provided observations. NOLAM (Lamanna and Serafini, 2024) can learn lifted action models even from noisy trajectories. LOCM (Cresswell and Gregory, 2011) and LOCM2 (Cresswell, McCluskey, and West, 2013), and its extension to learn action costs (Gregory and Lindsay, 2016) analyze only the sequences of actions in the given trajectories, ignoring any information about the states between them. Based on less structured input knowledge, Framer (Lindsay et al., 2017) is an approach for learning planning domain models from natural language descriptions of activity sequences.

The Safe Action Model Learning (SAM) learning algorithms (Stern and Juba, 2017; Mordoch, Juba, and Stern, 2023; Juba, Le, and Stern, 2021; Juba and Stern, 2022; Le, Juba, and Stern, 2024; Mordoch et al., 2024) are a family of action model learning algorithms that return action models that guarantee that plans generated with them are applicable in the actual action model. An action model having this property is referred to as *safe action model*. The SAM family of algorithms addresses learning safe action models under different settings: the action models are lifted (Juba, Le, and Stern, 2021), with stochastic (Juba and Stern, 2022) or conditional effects (Mordoch et al., 2024), and even action

models containing numeric preconditions and effects (Mordoch, Juba, and Stern, 2023). Le, Juba, and Stern (2024) extended their approach to support learning safe action models in a partially observable environment. ESAM (Juba, Le, and Stern, 2021) is an extension of the SAM algorithm that learns action models from domains in which there is ambiguity regarding the mapping of objects to parameters. To resolve this ambiguity, ESAM outputs a model with additional *proxy actions* that impose additional preconditions and parameter changes on actions in which such ambiguity cannot be resolved.

LatPlan (Asai and Fukunaga, 2018) and ROSAME-I (Xi, Gould, and Thiébaux, 2024) are conceptually different since they learn propositional action models from trajectories where the states in the given trajectories are given as images, as opposed to a conjunction of fluents. [Yarin: Rewrote and added part on LatPlan here:] LatPlan is a fully unsupervised system that uses a variational autoencoder as a differentiable approximation to convert image states into discrete propositional symbols, enabling classical planning in a learned latent space. ROSAME-I, in contrast, assumes a predefined set of propositions and action signatures. It simultaneously trains a classifier to identify propositions from images and learns a lifted, first-order action model over the given symbolic vocabulary.

3 Problem Setting and Syntactic Similarity

In this work we consider the following domain model learning setup. An agent is acting in an environment that we assume can be represented as a classical planning domain denoted by M^*. The agent's actions are recorded in a set of trajectories, which are sequences of observations and actions. The observations and actions in the trajectories are given in a specific representation, which we refer to as the *input representation*. A domain model learning algorithm is given this set of trajectories, and is expected to output a domain model in classical planning. That is, this domain model can be given as input to a classical planner, and together with an appropriate problem description, the planner can generate plans with it. [Gregor: Maybe add: These plans then should be applicable in the environment M^* and should lead to the goal set in the problem description.][Roni: I intentionally did not add this, as it is not always the case and we talk later about cases where we need a "decoder" to translate the plans generated by the planner and the thing that can be executed by the agent in the environment.]

The core question is then: How good is the model learned by the domain model learning algorithm? Many prior works evaluated their domain model learning algorithms by comparing the learned domain model to a *reference domain model* (or *ground truth model*). The reference domain model is assumed to be correct, i.e., equivalent to M^*, and has been used to evaluate the learned domain model in terms of its *syntactic similarity* to it (Aineto, Celorrio, and Onaindia, 2019; Mordoch et al., 2023; Xi, Gould, and Thiébaux, 2024; Oswald et al., 2024). These syntactic similarity metrics typically compare the intersection or difference of the predicates in the actions' preconditions and effects between the learned and reference domain models. We define these common metrics below. Let M and M^* be the evaluated and reference domain models, respectively, and let a be an action. We denote by $pre_M(a)$ the preconditions of action a according to domain M.

[Christian: "parameter-bound literals" needs to be defined. Also, at this point, it's not clear what the model learner is given. Are we using the same types? Same objects? Same action signatures? What about same fluent names with different number of parameters? I guess all these questions are why we need new metrics beyond the syntactic ones, but some of the assumptions that go into using these (syntactic) metrics would be useful to stipulate.] [Roni: Hmm. You're right there's some blunder in the text above. I removed "parameter-bound literal" altogether and stayed with simply literals.]

- True Positives: $\text{TP}_{pre}(a) = |(pre_M(a) \cap pre^*_M(a)|$
- False Positives: $\text{FP}_{pre}(a) = |((pre_M(a) \setminus pre^*_M(a)|$
- True Negatives: $\text{TN}_{pre}(a) = |L \setminus (pre_M(a) \cup pre^*_M(a))|$
- False Negatives: $\text{FN}_{pre}(a) = |(pre^*_M(a) \setminus pre_M(a))|$

The following standard metrics from statistical analysis can then be computed based on these values for each action:

- **Syntactic Precision**: $P_{pre}(a) = \frac{TP(a)}{TP(a)+FP(a)}$
- **Syntactic Recall**: $R_{pre}(a) \frac{TP(a)}{TP(a)+FN(a)}$

Other metrics, such as Accuracy and F1-score, can also be computed based on these values. To obtain an overall precision and recall for preconditions of the entire domain model, one can compute the average of the precision and recall values for all actions: $P_{\text{avg}} = \frac{1}{|A|} \sum_{a \in A} P(a)$ and $R_{\text{avg}} = \frac{1}{|A|} \sum_{a \in A} R(a)$, where $P(a)$ and $R(a)$ are the precision and recall of action a, respectively. The same metrics can be defined for the effects of actions, with the only difference being that the literals in the effects are used instead of those in the preconditions.

[Roni: TODO: Add an example] [Gregor: Proposed this "pathological" example] [Roni: Not sure about this. It is just an example, not a "corner case".] As an example, consider a delivery scenario. The true model M^* has predicates at, in, and $contains$ to describe that a truck or package is at a location, a package is in a truck, and a truck contains a package. The action unload has three parameters ℓ, t, and p (location, truck, and package). In M^* its preconditions are $at(\ell, t)$ and $in(p, t)$, while its effects are $\neg in(p, t)$, $\neg contains(t, p)$, and $at(p, t)$. A learned model could conceivably have the same effects, but the preconditions $at(\ell, t)$ and $contains(t, p)$. It would have syntactic recall and precision for the effects of 1, but for the preconditions it would be $\frac{1}{2}$.

In a slightly different fashion, Chrpa et al. (2023) proposed an approach to assess the edit distance of the learned domain model with regard to the reference one. Low distance values indicate models that are syntactically close to each other, and if two models are syntactically identical (i.e., edit distance of zero), then they are said to be *strongly equivalent* (Chrpa et al., 2023).

The above approach to evaluating domain models has several limitations. First, it relies on the existence of a reference domain model, which may not be available in practice. In fact, this is usually the case in new applications of automated planning to real-world problems. In such cases, there may be a number of alternative models, but not a specific reference one. This is the setting, for instance, of the International Competition on Knowledge Engineering for Planning and Scheduling (ICKEPS) (Chrpa et al., 2017). [Gregor: One could distinguish two cases here: (1) the practical application case, i.e. where we want to use domain learning in a real environment/industrial setting. Or (2) where we only want to evaluate the methods w.r.t. their usability. In the latter case, we could argue that we have "exhaustively" studied the setting we want to evaluate the model learner on and could thus have a planning model that models the situation – but still in this case, the issue of multiple equivalent models remains.]

[Gregor: Maybe add a new second reason here to prepare the argumentation for the next one:
Second, there might be multiple models that are fully identical for all practical purposes. That is, we will be able to find exactly the same set of plans using all of these models. A supposed ground truth model would then be an arbitrary pick out of these equivalent models. In evaluation would reward a learner if it can reproduce the one arbitrarily chosen ground truth model over any other equivalent model. Considering the above transport example, it is equivalent to use $in(p, t)$ or $contains(t, p)$ as a precondition – but using syntactic similarities forces the learner to have a bias for one of the two option, without being able to infer which bias is correct out of the input trajectories.]

Second, the reference model is assumed to be of the best possible quality. This is in itself complicated by the fact that it is very challenging to assess the quality of a model (McCluskey, Vaquero, and Vallati, 2017). Besides this, the use of a reference model biases the assessment towards a single specific encoding – while many others of similar quality could potentially exist. Third, it is not clear that the syntactic similarity of the learned model to a reference model is a good indicator of how *useful* a learned domain model is. This limitation has been observed in prior works (Aineto, Celorrio, and Onaindia, 2019; Juba, Le, and Stern, 2021; Mordoch et al., 2024). Next, we discuss what constitutes a useful domain model and how to evaluate it without the need for a reference domain model.

4 Metrics for Evaluating Domain Models

We can consider two main dimensions to optimize the learning process and to evaluate the learned model:

- **Predictive power.** Aim to learn a domain model that allows predicting the applicability and outcome of actions in the environment.
- **Problem-solving ability.** Aim to learn a domain model that enables the generation of applicable plans in the environment within reasonable resources bounds.

A key insight to consider is that these dimensions are not necessarily aligned with the syntactic similarity metrics defined above, even if a reference model exists that enables computing it. A domain model may be very *predictive* of behavior of the agent in the reference domain model, yet very different from a reference model since it encodes mutex conditions that the reference model did not bother to define. [Roni: Is this clear?][Yarin: Yes but, how is this instead: A domain model can be highly *predictive* of an agent's behavior in the reference domain, even if it differs significantly from the reference model. This can happen when the learned model encodes mutex (mutual exclusion) conditions that the reference model does not explicitly define.] [Gregor: maybe: "does only define implicitly"?] [Gregor: I've also put this issue already in the example above. Is this more illustrative?] Similarly, a learned domain model may be very *similar syntactically* to a reference domain model but not effective in solving problems from the corresponding real-world environment, e.g., adding redundant preconditions or missing a crucial effect (Vallati and Chrpa, 2019). [Gregor: or missing a crucial precondition: e.g. can only open a door if I have key.] In contrast, a domain model may be very different from a reference domain model, e.g., adding many redundant preconditions to some actions, yet very *effective* in solving problems in the application domain since these redundant preconditions are often true in the real world.

The two dimensions described here — predictive power and problem-solving ability — are also not necessarily aligned. Indeed, a domain model may be very *predictive* of behavior of the agent in the reference domain model, yet too complex to be used for solving problems by any existing planning engine. Assume, for example, that a learned domain model that has many copies of the same action in the reference domain model, each with different parameters and preconditions, in order to capture different aspects of that action's behavior. While this may be useful for predicting the applicability of actions in the reference domain model, it may hinder the ability of a planning engine to find plans for problems in the application domain with this model, due to its complexity (e.g., large branching factor). Therefore, different metrics are required to evaluate these two dimensions. We describe such metrics below.

4.1 Predictive Power Metrics

While the syntactic similarity metrics are easy to compute, they do not accurately reflect the *predictive power* of the learned model. Predictive power metrics, referred to sometimes as *semantic* domain model metrics (Aineto, Celorrio, and Onaindia, 2019; Mordoch et al., 2024; Le, Juba, and Stern, 2024), are based on the idea that a learned model should be able to predict the applicability of actions and their effects in the environment.

[Christian: Up until this point, I thought it was headed towards the predictive power of explaining the trajectory data. Given that it's about action applicability, this needs to be clarified much earlier on (e.g., intro). Also means that we don't have a notion of how well the learned model predicts the trajectories?]

We define two types of predictive power metrics: *action applicability* metrics and *predicted effects* metrics. The former measures the ability of the learned model to predict

whether an action is applicable in a given state, while the latter measures the ability of the learned model to predict the effects of an action in a given state. Unlike the syntactic similarity metrics, which are computed based on the action model itself, the predictive power metrics require a dataset of states that we denote by S. This dataset is intended to represent the distribution of states of interest in the domain. One way to create this dataset is by running a planner on a set of test problems with the real action model.[Gregor: Or to actually observe states that appear in the actual environment during execution.]

[Christian: We still need the original domain in order to compute the predictive power. Some of the criticisms lofted at the existing metrics may need to be scaled back because of this – there's no way to know if the predicted actions that are applicable are the right ones, without having a reference model.]

[Roni: TODO: Clarify this does not consider complexity of the domain (which will be addressed in the next metric)]

[Gregor: Issue: what happens if we either cannot get the states themselves, i.e., no symbolic description or if the state description of the actual model differs from the one of the learned model? (say: predicates are named differently?)
Can we be more radical here? Maybe it is sufficient to have plans and non-plans for the real model? Then we could check whether these plans are executable in the learned model and whether the non-plans are not. One issue: this mixes applicability and effects.]

[Christian: Ya, I think many of the definitions start to fall apart if we don't have several assumptions down – same objects, same types, same fluent/action symbols with their signatures, etc.]

Predicted applicability For a domain model M and action a, we denote by $app_M(a, S)$ the set of states in S in which a is applicable according to M. [Gregor: This is easy to compute for STRIPS/SAS+, but hard if you have things like disjunctive preconditions. It should be #P-hard] [Roni: Hmm. I don't see what this is not linear in the size of S and the preconditions of a in M: we iterate over every state in S and then check whether a is applicable in it according to M. I guess if we have some universals in the preconditions this might be harder?] Based on this notation, we define the following predicted action applicability metric as follows for some action a:

- $\text{TP}_{app}(a) = |app_M(a, S) \cap app^*_M(a, S)|$
- $\text{FP}_{app}(a) = |app_M(a, S) \setminus app^*_M(a, S)|$
- $\text{TN}_{app}(a) = |S \setminus (app_M(a, S) \cup app^*_M(a, S))|$
- $\text{FN}_{app}(a) = |app^*_M(a, S) \setminus app_M(a, S)|$

In words, $\text{TP}_{app}(a)$ is the number of states in S where a is applicable according to both the learned model and the real model, $\text{FP}_{app}(a)$ is the number of states in S where a is applicable according to the learned model but not the real model, $\text{TN}_{app}(a)$ is the number of states in S where a is not applicable according to both models, and $\text{FN}_{app}(a)$ is the number of states in S where a is applicable according to the real model but not the learned model. From these metrics, one can compute the precision and recall of the learned model for action applicability as follows,

$$P_{app}(a) = \frac{TP_{app}(a)}{TP_{app}(a) + FP_{app}(a)} \quad (1)$$

$$R_{app}(a) = \frac{TP_{app}(a)}{TP_{app}(a) + FN_{app}(a)} \quad (2)$$

and the overall action applicability, precision, and recall by averaging over all actions.

Predicted effects For domain M, action a, and state s, we denote by $a_M(s)$ the state resulting from applying a in s according to M. Based on this, we define the following predicted action effect metrics for some action a and state s, as follows:

- $\text{TP}_{\mathit{eff}}(s, a) = |(a_M(s) \setminus s) \cap (a^*_M(s) \setminus s)|$
- $\text{FP}_{\mathit{eff}}(s, a) = |(a_M(s) \setminus s) \setminus a^*_M(s)|$
- $\text{TN}_{\mathit{eff}}(s, a) = |s \cap a_M(s) \cap a^*_M(s)|$
- $\text{FN}_{\mathit{eff}}(s, a) = |(a_M(s) \cap s) \setminus a^*_M(s)|$ [Gregor: there is a start too many (this is always $\emptyset$) I assume it should be $(a_M(s) \cap s) \setminus (a^*_M(s) \cap s)$ – the facts that are not changed by a_M, but not counting the ones changes by M^*_M][Roni: Good catch! I fixed it in a slightly different way, let me know if you agree or not.][Christian: The TN and FN aren't looking correct...but after spending some minutes with it, I think it is ;). Should state that a "negative" here indicates a literal that remains unchanged.]

For the purpose of the above computation, a state includes all the literals true in it, i.e., both positive propositions and negative ones. One can aggregate the above metrics over all states and actions, [Gregor: This might be computationally a problem – so we sample?!][Roni: Hmm. I am not sure. Is it worse than low-order polynomial in S and number of actions?] compute the overall values, and compute aggregated precision and recall values accordingly. The difference between the syntactic similarity of effects and the predicted effects metrics is subtle. It manifests when processing observations in which some literal ℓ is observed in the state before an action a, it is not an [Gregor: posisitve i.e. adding?] effect according to the learned model, but it is an effect according to the real model. This will count as a false positive in the syntactic similarity metrics yet not count as a false positive in the predicted effects metrics.

[Roni: Add a concrete example of how to compute these metrics] [Yarin: I can add an example after we verify the different formulas, example for each or only for Predicted effects ?]

4.2 Problem-Solving Metrics

Neither the syntactic similarity metrics nor the predictive power metrics are sufficient for evaluating the operationality (McCluskey, Vaquero, and Vallati, 2017) of the learned domain model, i.e. its ability to solve problems. It is well-known that even small syntactical changes in domain models can result in significant performance gaps (Vallati and Chrpa, 2019; Vallati et al., 2021).

We propose two metrics that are designed to address this issue: *solving ratio* and *false plans ratio*. Both metrics are defined with respect to a set of problems P in the environment M^*. The solving ratio metric is defined as the ratio of problems in P that can be solved with the learned model by a given planning within a fixed limit on the available computational resources — CPU runtime and memory. We say that a problem is solved if a plan is found within the limited computational resources, and that plan is valid according to the environment. The false plans ratio metric is defined as the ratio of problems in P that are solved by the learned model but are not valid according to the environment. This reflects the reliability of using plans with the learned model.

The straightforward nature of the above metrics is not without limitations. The ability to solve a problem with a given domain depends on external factors such as the set of test problems, the planner used, the runtime and memory budget allowed for it to run, and the computer and OS that executed the planner. Ideally, the above metrics would be run on a diverse set of test problems, planners, resource limits, and computing machines. In practice, this might be difficult to implement, but one is advised to at least run all evaluated domains on the same setup and provide an appropriate disclaimer to the concluded result.

[Roni: TODO: Add metric on the runtime of solving the problem.] [Roni: TODO: Discussion: what about a metric on how many real plans can be validated by the learned model.][Mauro: Yes! validation capability of the learned model with regards to plans generated with the reference model!] [Roni: TODO: Maybe: add some unsolvability detection metrics?] [Christian: In our work on aligning (which is very closely related, it seems!), we computed both (1) if the plans found using P and M validate on M* and (2) if the plans found using P and M* validate on M. You need assumptions on the action names+parameters, but then can test (via VAL) both ways.]

Note that all the metrics described in this section do not require a reference model, but instead are with respect to the actual environment. Thus, they can be used to compare two models directly, providing statistics on which one is better according to different aspects of the environment. [Christian: I think the above is misleading. We need to have states (fluents matching the learned model), action applicability (for predictive power), etc, etc. We do need a reference model! What's changed is that we aren't using syntactic comparisons anymore, but the M* model is certainly still there.]

5 Bridging Representation Gaps

Some domain model learning algorithms output a classical planning domain model that uses a different representation than the input representation. Such a model can still be used by a planner, yet its input and output may be incompatible with that of the reference domain model. In some cases, these representation differences are minimal, e.g., using different ordering of the parameters or object types. In other cases, the differences are more significant. For example, the ESAM algorithm (Juba, Le, and Stern, 2021) outputs a domain model with additional *proxy actions* that are used to resolve ambiguities in the mapping of objects to parameters. [Roni: Pascal mentioned a paper that learned macro actions. TODO: add ref for it here (with relevant short text)] The proxy actions are not part of the original domain model, and they are not used in the reference domain model. [Roni: More examples?]

[Mauro: OpMaker2 generates classical planning models in OCL language McCluskey et al. (2010)]

[Yarin: added LOCM:] Similarly, the LOCM algorithm (Cresswell, McCluskey, and West, 2013) induces finite-state machines solely from action sequences-with no access to intermediate state information or predicate names-and outputs action schemas using *proxy predicates*. These proxy predicates are not part of the original domain but are introduced as a mechanism to infer the semantics of the original predicates.

The syntactic similarity metrics are not relevant in such cases, where the representation is intentionally different from the reference model. Next, we describe how to apply the other metrics – predictive power and problem-solving – in such cases. Note that we limit the discussion to the case where the learned domain is still a planning domain, as opposed to images or other more involved representations.

5.1 The Encoder-Decoder Mechanism

Let R_{M^*} denote the input representation and let R_M denote the representation of the learned domain model. The given trajectories are given in R_{M^*}, while the output of a planner that uses the learned domain model is in R_M. To allow the comparison of the learned domain model with the reference domain model, we need to bridge the gap between these two representations. To this end, we require every domain model learning algorithm to output a representation *encoder* and a representation *decoder* in addition to the learned model. We describe these two components below. The *encoder* transforms states and actions in R_{M^*} to corresponding states and actions in the learned domain representation R_M. The *decoder* performs the reverse transformation, mapping states and actions in R_M to corresponding states and actions in R^*_M.

Formally, let A and A_M be the set of actions in the real and learned domain models, respectively, and let S and S_M be the set of states in the real and learned domain models, respectively. The power set is denoted by $P(X)$, which is the set of all subsets of X. The encoder is required to implement the following set of functions:

- $encodeAction : S \times A \to P(A_M)$, returns the set of actions in M that represent the application of a[Yarin: $a \in A$] in a given state s[Yarin: $s \in S$].
- $encodeState : S \to S_M$, returns the state in M that represents the state s in M^*. [Yarin: returns the state $s \in S$ as its representation in S_M]

The decoder is required to implement the following functions:

- $decodeAction : A_M \to A$, returns the action in M^* corresponding to the action in M. [Yarin: returns the action $a \in A_M$ as its representation in A]

- $decodeState : S_M \to S$, returns the state in the input representation (S) that represents a given state in S_M. [Yarin: returns the state $s \in S_M$ as its representation in S]

5.2 Predictive Power Metrics with an Encoder-Decoder

For the predicted applicability metric, we modify the way $app_M(a, S)$ is computed. The main difference is that a single action in the reference domain model may be represented by multiple actions in the learned model. Thus, we define $app_M(a, S)$ to be the number of states in S in which *there exists* an action a' in the learned model such that a' is applicable in the state in M that encodes s.

[Roni: TODO: Add a formal definition of the app_M method using the encoder and decoder methods]

Similarly, for the predicted effects metric, we need to decode the state resulting from $a_M(s)$ instead of using it as-is.

[Roni: TODO: Add a formal definition of the app_M method using the encoder and decoder methods]

5.3 Problem-Solving Metrics with an Encoder-Decoder

Adapting the problem-solving metrics to use the encoder-decoder mechanism is straightforward. The encoder is used to encode the problem from the input representation to the learned model representation. Then, the planner is run on the encoded problem with the learned model. Finally, the decoder is used to decode the solution plan from the learned model representation to the input representation. This allows checking if a plan has been found and if it is valid according to the reference domain model.

5.4 Case Studies

Next, we describe how the encoder-decoder mechanism can be implemented for several action model learning algorithms. [Yarin: Some action model learning algorithms operate directly in the original input representation R_{M^*}, without transforming the state or action spaces. For these algorithms, the learned model uses the same symbols and structure as the input data. As such, there is no need for additional encoding or decoding: the identity function suffices for both.]

SAM (Stern and Juba, 2017)[Yarin: or is ut juba2021safe ?]: [Roni: Either me or someone from my group will write this one] Encoder: identity Decoder: identity

FAMA (Aineto, Celorrio, and Onaindia, 2019) - Leonardo L.?: Encoder: identity Decoder: identity - maybe change to parameter names?

ESAM (Juba, Le, and Stern, 2021): [Roni: Either me or someone from my group will write this one] Encoder: identity for both state and actions Decoder: translate proxy actions to original actions

ROSAME-I (Xi, Gould, and Thiébaux, 2024): [Roni: Maybe Yarin or Argaman can help with this one] Encoder: encode to a numeric vector of ones and zeros Decoder: map action parameters

OBSERVER[Yarin: add cite]: [Roni: Anyone can help here?] Encoder: from binding to grounded Decoder: from grounded predicates and actions to the binding they represent

LOCM (Cresswell, McCluskey, and West, 2013):[Yarin: I will complete this] Encoder: identity for action name and proxy for state predicates Decoder: translate proxy predicates to original predicates

[Yarin: Add NOLAM too?] [Roni: Most important: anyone can help here?]

5.5 An Evaluation Paradigm

Using the above metrics and the encoder-decoder mechanism, we can define an evaluation paradigm for action model learning algorithms. The evaluation paradigm consists of the following steps: First, we generate a set of trajectories in the input representation for learning. This set of trajectories can be generated by running a planner on a set of problems in the reference domain model or via a random walk. Then, this set of trajectories is split in 80/20 ratio, using 80% of the trajectories for training and 20% for testing. The evaluated learning algorithm is given the training set of trajectories and is required to output a learned domain model, an encoder, and a decoder. [Roni: TODO for myself: Complete this subsection] This 80/20 split is repeated k times following a standard k-fold validation process.

[Christian: An eval-heavy ICAPS'er may take issue with some of the suggestions (random walks are hard to do right, the benchmark problem sets aren't IID, etc, etc. I'd say we're probably mostly fine (someone can always find some issue if they try hard enough), but we *should* comment on the disconnect of trace -vs- state. Many approaches require traces, but the evals require state sets. Taking the latter to just be states along traces is a deliberate choice that warrants discussion.]

6 Benchmarks

[TODO: Leonardo is in charge of this section.]

[TODO: Describe the benchmark suite and how to use it.]

[TODO: If we have time: show results for at least some of the algorithms on the benchmark suite.]

7 Discussion

[Roni: For all: please read and let me know what you think by adding comments in the text (e.g., add text like this "[[YourName: bla bla bla"]]) or editting.] The proposed evaluation paradigm is designed to be flexible and extensible. In particular, it can be extended to support other types of representations, such as images or other types of data. Another possible extension is to allow the action model learning algorithm to output a domain model that uses a more general type of planning. For example, the action model learning algorithm may output a domain model that uses a probabilistic (Xi, Gould, and Thiébaux, 2024) or partially observable representation (Le, Juba, and Stern, 2024). Adapting the predictive power metrics to such cases may not be trivial, but the problem-solving metrics can be adapted in a straightforward manner.

Going beyond discrete representations is also possible. The predicted applicability metrics can be computed as-is,

considering both numeric and discrete preconditions. More involved is the predicted effects metric, which may require defining some loss function for the numeric effects, as learning exactly the same numeric effects as in the reference domain model seems unlikely. Note that in all of these variants and generalizations, the encoder-decoder mechanism allows seamless use of the problem-solving metrics.

[TODO: Maybe talk about metrics for online learning] The metrics and evaluation paradigm outlined in this paper are for *offline* learning of action models, where the learning algorithm is given a set of trajectories and is required to output a domain model. *Online learning* of action models have also been studied in the literature (Lamanna et al., 2021; Sreedharan and Katz, 2023; Benyamin et al., 2025; Ng and Petrick, 2019; Chitnis et al., 2021; Verma, Karia, and Srivastava, 2023; Karia et al., 2023; Jin et al., 2022)[TODO: More citations?][Yarin: added, to many?], where the learning algorithm is required to learn a domain model by actively interacting with the environment. Similar to other online tasks, one may distinguish between the *cumulative regret* of the learning process, which can be the number of actions performed for learning or the number of problems failed to solve while collecting observations. Specific metrics and evaluation for online learning of action models, however, is beyond the scope of this work.

Finally, models can also be compared not only in terms of their characteristics, but also in terms of the processes used to learn or generate them (Vallati and McCluskey, 2021), a perspective that is commonly used for assessing conceptual models.

7.1 Relation to Model Reconciliation

[Roni: Pascal and Mauro: maybe put here is a discussion on what is model reconciliation and how it relates to our work, and existing model reconciliation metrics?]

7.2 Relation to Model Repair

[Roni: Pascal and Mauro: same for model repair?]

8 Conclusion Future Work

[Roni: I'll fix this text later] We presented a new paradigm for evaluating action model learning algorithms across different representations. In this paradigm, each action model learning algorithm is required to output an action model and an encoder-decoder pair. The encoder-decoder pair is used to bridge representation gaps and enable measuring the problem-solving capabilities of the learned action models. We proposed an evaluation scheme that leverages the encoder-decoder pair to systematically compare learned action models and described several evaluation metrics. A benchmark suite was also provided to facilitate the evaluation of action model learning algorithms, based on existing domain models from the International Planning Competition (IPC). We demonstrated our evaluation paradigm by applying it to several action model learning algorithms, including SAM, ESAM, FAMA, and ROSAME.

References

Aineto, D.; Celorrio, S. J.; and Onaindia, E. 2019. Learning action models with minimal observability. *Artificial Intelligence* 275:104–137.

Amir, E., and Chang, A. 2008. Learning partially observable deterministic action models. *Journal of Artificial Intelligence Research* 33:349–402.

Asai, M., and Fukunaga, A. 2018. Classical planning in deep latent space: Bridging the subsymbolic-symbolic boundary. In *Proceedings of the aaai conference on artificial intelligence*, volume 32, 6094–6101.

Benyamin, Y.; Mordoch, A.; Shperberg, S. S.; and Stern, R. 2025. Integrating reinforcement learning, action model learning, and numeric planning for tackling complex tasks.

Chitnis, R.; Silver, T.; Tenenbaum, J. B.; Kaelbling, L. P.; and Lozano-Pérez, T. 2021. Glib: Efficient exploration for relational model-based reinforcement learning via goal-literal babbling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 11782–11791.

Chrpa, L.; McCluskey, T. L.; Vallati, M.; and Vaquero, T. 2017. The fifth international competition on knowledge engineering for planning and scheduling: Summary and trends. *AI Mag.* 38(1):104–106.

Chrpa, L.; Dodaro, C.; Maratea, M.; Mochi, M.; and Vallati, M. 2023. Comparing planning domain models using answer set programming. In *European Conference on Logics in Artificial Intelligence*, 227–242.

Cresswell, S., and Gregory, P. 2011. Generalised domain model acquisition from action traces. In *International Conference on Automated Planning and Scheduling (ICAPS)*, 42–49.

Cresswell, S. N.; McCluskey, T. L.; and West, M. M. 2013. Acquiring planning domain models using locm. *The Knowledge Engineering Review* 28(2):195–213.

Gregory, P., and Lindsay, A. 2016. Domain model acquisition in domains with action costs. In *International Conference on Automated Planning and Scheduling (ICAPS)*, 149–157.

Gösgens, J.; Jansen, N.; and Geffner, H. 2025. Learning lifted STRIPS models from action traces alone: A simple, general, and scalable solution. In *35th International Conference on Automated Planning and Scheduling (ICAPS 2025)*.

Haslum, P.; Lipovetzky, N.; Magazzeni, D.; Muise, C.; Brachman, R.; Rossi, F.; and Stone, P. 2019. *An introduction to the planning domain definition language*.

Helmert, M. 2006. The fast downward planning system. *Journal of Artificial Intelligence Research* 26:191–246.

Jin, M.; Ma, Z.; Jin, K.; Zhuo, H. H.; Chen, C.; and Yu, C. 2022. Creativity of AI: Automatic symbolic option discovery for facilitating deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 7042–7050.

Juba, B., and Stern, R. 2022. Learning probably approximately complete and safe action models for stochastic worlds. In *AAAI*, 9795–9804.

Juba, B.; Le, H. S.; and Stern, R. 2021. Safe learning of lifted action models. In *International Conference on Principles of Knowledge Representation and Reasoning (KR)*, 379–389.

Karia, R.; Verma, P.; Vipat, G.; and Srivastava, S. 2023. Epistemic exploration for generalizable planning and learning in non-stationary stochastic settings. In *NeurIPS 2023 Workshop on Generalization in Planning*.

Lamanna, L., and Serafini, L. 2024. Action model learning from noisy traces: a probabilistic approach. In *ICAPS*, 342–350. AAAI Press.

Lamanna, L.; Saetti, A.; Serafini, L.; Gerevini, A.; and Traverso, P. 2021. Online learning of action models for pddl planning. In *IJCAI*, 4112–4118.

Lamanna, L.; Serafini, L.; Saetti, A.; Gerevini, A. E.; and Traverso, P. 2025. Lifted action models learning from partial traces. *Artificial Intelligence* 339:104256.

Le, H. S.; Juba, B.; and Stern, R. 2024. Learning safe action models with partial observability. In *AAAI Conference on Artificial Intelligence*, volume 38, 20159–20167.

Lindsay, A.; Read, J.; Ferreira, J. F.; Hayton, T.; Porteous, J.; and Gregory, P. 2017. Framer: Planning models from natural language action descriptions. In *International Conference on Automated Planning and Scheduling (ICAPS)*.

McCluskey, T. L.; Cresswell, S. N.; Richardson, N. E.; and West, M. M. 2010. Action knowledge acquisition with opmaker2. In *Agents and Artificial Intelligence*, 137–150.

McCluskey, T. L.; Vaquero, T. S.; and Vallati, M. 2017. Engineering knowledge for automated planning: Towards a notion of quality. In *Proceedings of the Knowledge Capture Conference, K-CAP*, 14:1–14:8.

Mordoch, A.; Stern, R.; Scala, E.; and Juba, B. 2023. Safe learning of pddl domains with conditional effects. In *Reliable Data-Driven Planning and Scheduling (RDDP) workshop in ICAPS*.

Mordoch, A.; Scala, E.; Stern, R.; and Juba, B. 2024. Safe learning of pddl domains with conditional effects. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 34, 387–395.

Mordoch, A.; Juba, B.; and Stern, R. 2023. Learning safe numeric action models. In *AAAI Conference on Artificial Intelligence*, 12079–12086.

Ng, J. H. A., and Petrick, R. P. 2019. Incremental learning of planning actions in model-based reinforcement learning. In *IJCAI*, 3195–3201.

Oswald, J.; Srinivas, K.; Kokel, H.; Lee, J.; Katz, M.; and Sohrabi, S. 2024. Large language models as planning domain generators. In *Proceedings of the 34th International Conference on Automated Planning and Scheduling (ICAPS 2024)*, 423–431. AAAI Press.

Sreedharan, S., and Katz, M. 2023. Optimistic exploration in reinforcement learning using symbolic model estimates. *Advances in Neural Information Processing Systems* 36:34519–34535.

Stern, R., and Juba, B. 2017. Efficient, safe, and probably approximately complete learning of action models. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 4405–4411.

Vallati, M., and Chrpa, L. 2019. On the robustness of domain-independent planning engines: The impact of poorly-engineered knowledge. In *Proceedings of the 10th International Conference on Knowledge Capture, K-CAP*, 197–204.

Vallati, M., and McCluskey, T. L. 2021. A quality framework for automated planning knowledge models.

Vallati, M.; Chrpa, L.; McCluskey, T. L.; and Hutter, F. 2021. On the importance of domain model configuration for automated planning engines. *Journal of Automated Reasoning* 65(6):727–773.

Verma, P.; Karia, R.; and Srivastava, S. 2023. Autonomous capability assessment of sequential decision-making systems in stochastic settings. *Advances in Neural Information Processing Systems* 36:54727–54739.

Xi, K.; Gould, S.; and Thiébaux, S. 2024. Neuro-symbolic learning of lifted action models from visual traces. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 34, 653–662.

Yang, Q.; Wu, K.; and Jiang, Y. 2007. Learning action models from plan examples using weighted max-sat. *Artificial Intelligence* 171(2-3):107–143.

Zhuo, H. H.; Yang, Q.; Hu, D. H.; and Li, L. 2010. Learning complex action models with quantifiers and logical implications. *Artificial Intelligence* 174(18):1540–1569.