

# Top-Down Synthesis for Library Learning

MATTHEW BOWERS, Massachusetts Institute of Technology, USA

THEO X. OLAUSSON, Massachusetts Institute of Technology, USA

CATHERINE WONG, Massachusetts Institute of Technology, USA

GABRIEL GRAND, Massachusetts Institute of Technology, USA

JOSHUA B. TENENBAUM, Massachusetts Institute of Technology, USA

KEVIN ELLIS, Cornell University, USA

ARMANDO SOLAR-LEZAMA, Massachusetts Institute of Technology, USA

This paper introduces *corpus-guided top-down synthesis* as a mechanism for synthesizing library functions that capture common functionality from a corpus of programs in a domain specific language (DSL). The algorithm builds abstractions directly from initial DSL primitives, using syntactic pattern matching of intermediate abstractions to intelligently prune the search space and guide the algorithm towards abstractions that maximally capture shared structures in the corpus. We present an implementation of the approach in a tool called STITCH and evaluate it against the state-of-the-art deductive library learning algorithm from DreamCoder. Our evaluation shows that STITCH is 3-4 orders of magnitude faster and uses 2 orders of magnitude less memory while maintaining comparable or better library quality (as measured by compressivity). We also demonstrate STITCH’s scalability on corpora containing hundreds of complex programs that are intractable with prior deductive approaches and show empirically that it is robust to terminating the search procedure early—further allowing it to scale to challenging datasets by means of early stopping.

Additional Key Words and Phrases: Program Synthesis, Library Learning, Abstraction Learning

## ACM Reference Format:

Matthew Bowers, Theo X. Olausson, Catherine Wong, Gabriel Grand, Joshua B. Tenenbaum, Kevin Ellis, and Armando Solar-Lezama. 2023. Top-Down Synthesis for Library Learning. *Proc. ACM Program. Lang.* 7, POPL, Article 41 (January 2023), 42 pages. <https://doi.org/10.1145/3571234>

## 1 INTRODUCTION

One way programmers manage complexity is by hiding functionality behind functional abstractions. For example, consider the graphics programs at the bottom of Fig. 1A. Each uses a generic set of drawing primitives and renders a technical schematic of a hardware component, shown to the left of each program. Faced with the task of writing *more* of these rendering programs, an experienced human programmer likely would not continue using these low-level primitives. Instead, they would introduce new functional abstractions, like the one at the bottom of Fig. 1B which renders a regular polygon given a size and number of sides. Useful abstractions like these allow more concise and

---

Authors’ addresses: Matthew Bowers, Massachusetts Institute of Technology, Cambridge, MA, USA, [mlbowers@mit.edu](mailto:mlbowers@mit.edu); Theo X. Olausson, Massachusetts Institute of Technology, Cambridge, MA, USA, [theoxo@mit.edu](mailto:theoxo@mit.edu); Catherine Wong, Massachusetts Institute of Technology, Cambridge, MA, USA, [catwong@mit.edu](mailto:catwong@mit.edu); Gabriel Grand, Massachusetts Institute of Technology, Cambridge, MA, USA, [grandg@mit.edu](mailto:grandg@mit.edu); Joshua B. Tenenbaum, Massachusetts Institute of Technology, Cambridge, MA, USA; Kevin Ellis, Cornell University, Ithaca, NY, USA; Armando Solar-Lezama, Massachusetts Institute of Technology, Cambridge, MA, USA.

---

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

© 2023 Copyright held by the owner/author(s).

2475-1421/2023/1-ART41

<https://doi.org/10.1145/3571234>

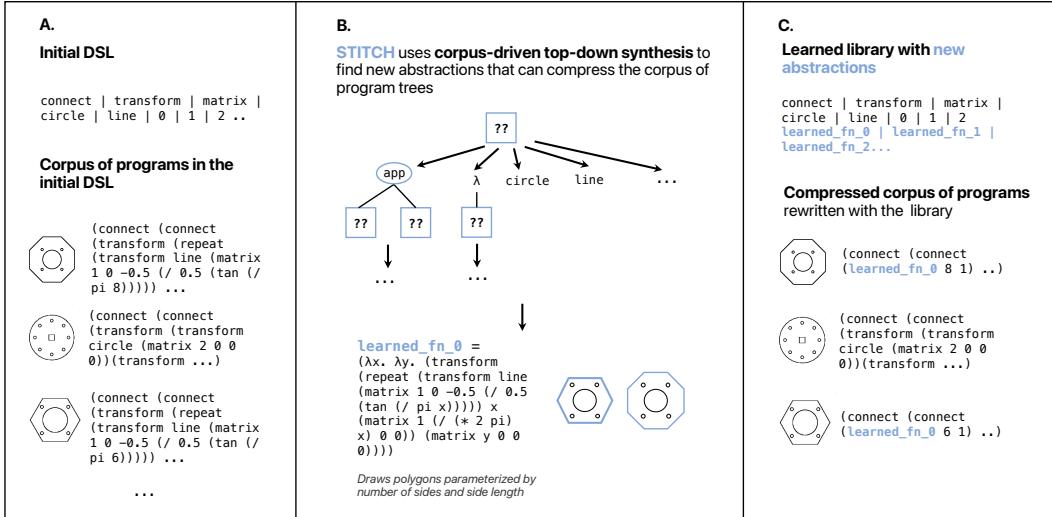


Fig. 1. (A) Given an initial DSL of lower-level primitives and a corpus of programs written in the initial DSL, (B) STITCH uses a fast and memory efficient *corpus-guided top-down search* algorithm to construct function abstractions which maximally capture shared structure across the corpus of programs. (C) STITCH automatically rewrites the initial corpus using the learned library functions.

legible programs to render the existing schematics. More importantly, well-written abstractions should generalize, making it easier to write new graphics programs for similar graphics tasks.

Recently, the *program synthesis* community has introduced new approaches that can mimic this process, automatically building a library of functional abstractions in order to tackle more complex synthesis problems [Dechter et al. 2013; Ellis et al. 2020, 2021; Lázaro-Gredilla et al. 2019; Shin et al. 2019]. One popular approach to library learning is to search for common tree fragments across a corpus of programs, which can be introduced as new abstractions [Dechter et al. 2013; Lázaro-Gredilla et al. 2019; Shin et al. 2019]. Ellis et al. 2021, however, introduces an algorithm that reasons about variable bindings to abstract out well-formed *functions* instead of just tree fragments. While it produces impressive results, the system in Ellis et al. 2021 takes a *deductive* approach to library learning that is difficult to scale to larger datasets of longer and deeper input programs. This approach is deductive in that it uses semantics-preserving rewrite rules to attempt to refactor existing programs to expose shared structure. This requires representing and evaluating an exponentially large space of proposed refactorings to identify common functionality across the corpus. Prior work, such as Ellis et al. 2020, 2021, approaches this challenge by combining a dynamic bottom-up approach to refactoring with version spaces to more efficiently search over the refactored programs. However, these deductive approaches face daunting memory and search requirements as the corpus scales in size and complexity.

This paper introduces an alternate approach to library learning, while preserving the focus on well-formed function abstractions from Ellis et al. 2021. Instead of taking a *deductive* approach based on refactoring the corpus with rewrite rules, we *directly synthesize abstractions*. We call this approach *corpus-guided top-down synthesis*, and it is based on the insight that when applied to the task of synthesizing abstractions, top-down search can be guided precisely towards discovering shared abstractions over a set of existing training programs. At every step of the search, syntactic comparisons between a partially constructed abstraction and the set of training programs can be

used to strongly constrain the search space and direct the search towards abstractions that capture the greatest degree of shared syntactic structure.

We implement this approach in **STITCH**, a corpus-guided top-down library learning tool written in Rust (Fig. 1). **STITCH** is open-source and offers both Python and Rust bindings; the code, bindings, and tutorial are available on the GitHub<sup>1</sup>. We evaluate **STITCH** through a series of experiments (Section 6), and find that: **STITCH** learns libraries of comparable quality to those found by the algorithm of Ellis et al. 2021 on their iterative bootstrapped library learning task, while being 3-4 orders of magnitude faster and using 2 orders of magnitude less memory (Section 6.1); **STITCH** learns high quality libraries within seconds to single-digit minutes when run on corpora containing a few hundred programs with mean lengths between 76–189 symbols (sourced from Wong et al. 2022), while even the simplest of these corpora lies beyond the reach of the algorithm of Ellis et al. 2021 (Section 6.2); and that **STITCH** degrades gracefully when resources are constrained (Section 6.3). We also perform ablation studies to expose the relative impact of different optimizations in **STITCH** (Section 6.4). Finally, we show that **STITCH** is complementary to deductive rewrite approaches (Section 6.5).

In summary, our paper makes the following contributions:

- (1) **Corpus-guided top-down synthesis (CTS)**: A novel, strongly-guided branch-and-bound algorithm for synthesizing function abstractions (Sec. 3).
- (2) **CTS for program compression**: An instantiation of the CTS framework for utility functions favoring abstractions that compress a corpus of programs (Sec. 4).
- (3) **STITCH**: A parallel Rust implementation of CTS for compression that achieves 3-4 orders of magnitude of speed and memory improvements over prior work, and the analysis of its performance, scaling, and optimizations through several experiments (Sec. 6).

## 2 OVERVIEW

In this section, we build intuition for the algorithmic insights that power **STITCH**. As a running example, we focus on learning a single abstraction from the following corpus of programs:

$$\begin{aligned} &\lambda x. + \ 3 \ (* \ (+ \ 2 \ 4) \ 2) \\ &\lambda xs. \text{map} \ (\lambda x. + \ 3 \ (* \ 4 \ (+ \ 3 \ x))) \ xs \\ &\lambda x. * \ 2 \ (+ \ 3 \ (* \ x \ (+ \ 2 \ 1))) \end{aligned} \tag{1}$$

The notion of what a *good* abstraction is depends on the application, so our algorithm is generic over the utility function that we seek to maximize. Following prior work [Dechter et al. 2013; Ellis et al. 2021; Shin et al. 2019] we focus on *compression* as a utility function: a good abstraction is one which minimizes the size of the corpus when rewritten with the abstraction. The utility function used by **STITCH** is detailed in Section 4 and corresponds exactly to the compression objective, but at a high level the function seeks to maximize the product of *the size of the abstraction* and the *number of locations* where the abstraction can be used. This product balances two key features of a highly compressive abstraction: the abstraction should be general enough that it applies in many locations, but specific enough that it captures a lot of structure at each location.

The optimal abstraction that maximizes our utility in this example is:

$$\text{fn}\theta = \lambda\alpha. \lambda\beta. (+ \ 3 \ (* \ \alpha \ \beta)) \tag{2}$$

---

<sup>1</sup><https://github.com/mlb2251/stitch>

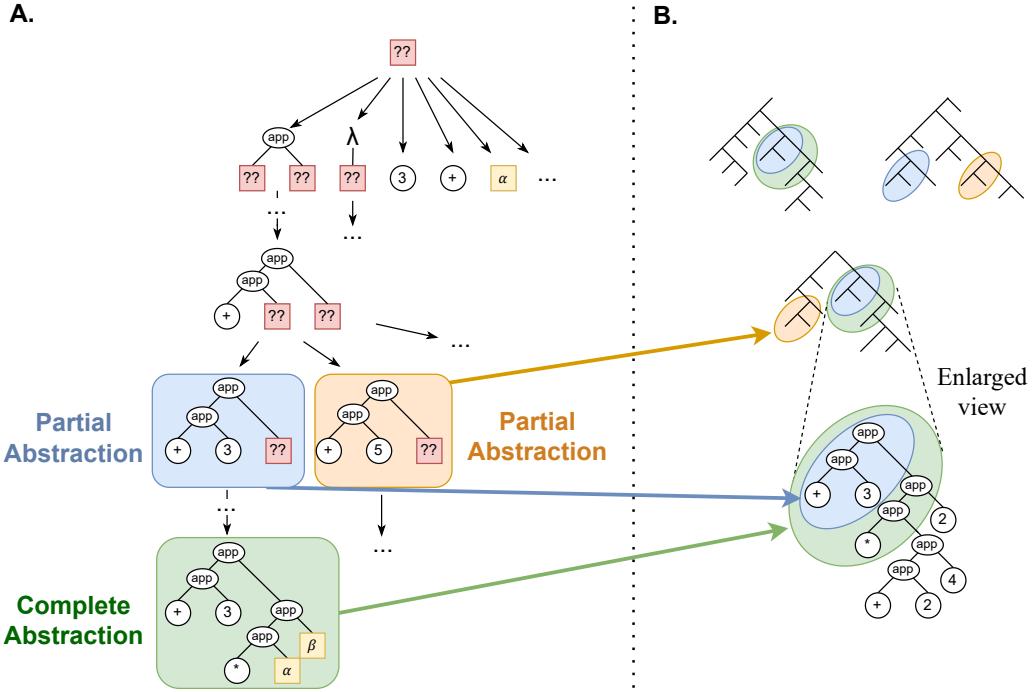


Fig. 2. (A) Schematic of *corpus-guided top-down search* (CTS). Partial abstractions can contain *holes* indicating unfinished subtrees, denoted ??, while complete abstractions do not contain holes. (B) Partial and complete abstractions *match* at locations in the corpus. The blue partial abstraction can be expanded into the green complete abstraction so the green abstraction matches in a *subset* of places that the blue abstraction does. For complete abstractions, *matching* indicates that the subtree can be rewritten to use the abstraction.

And the shared structure that this abstraction captures is highlighted in blue in Eq. (1). When rewritten to use this abstraction, the size of the resulting programs is minimized:

$$\begin{aligned}
 & \lambda x. \text{fn0 } (+ 2 4) 2 \\
 & \lambda xs. \text{map } (\lambda x. \text{fn0 } 4 (+ 3 x)) xs \\
 & \lambda x. * 2 (\text{fn0 } x (+ 2 1))
 \end{aligned} \tag{3}$$

STITCH synthesizes the optimal abstraction directly through *top-down search*. Top-down search methods construct program syntax trees by iteratively refining partially-completed programs that have unfinished *holes*, until a complete program that meets the specification is produced [Balog et al. 2016; Ellis et al. 2021; Feser et al. 2015; Nye et al. 2021; Polikarpova et al. 2016; Shah et al. 2020]. This kind of search is often made efficient by identifying branches of the search tree that can be pruned away because the algorithm can efficiently determine that none of the programs in that branch can be correct. We aim to apply a similar idea to the problem of synthesizing good abstractions; the idea is to explore the space of functions in the same top-down way, but in search of a function that maximizes the utility measure. The key observation is that this new objective affords even more aggressive pruning opportunities than the traditional correctness objective, allowing us to synthesize optimal abstractions very efficiently. We call this approach *corpus-guided top-down search* (CTS).

**Corpus-Guided Top-Down Search.** Like other top-down synthesis approaches, our algorithm explores the space of abstractions by repeatedly refining abstractions with holes, as illustrated in Fig. 2A. We call abstractions with holes *partial abstractions* in contrast to *complete abstractions* which have no holes. Our top-down algorithm searches over abstraction *bodies*, so to synthesize the optimal abstraction  $\text{fn}0$  in our running example, we synthesize its body:  $(+ 3 (* \alpha \beta))$ .

We say that a partial abstraction *matches* at a subtree in the corpus if there's a possible assignment to the holes and arguments that yields the subtree. For example, consider the subtree  $(+ 3 (* (+ 2 4) 2))$  from the first program in our running example, shown as a syntax tree at the bottom of figure 2B. The partial abstraction  $(+ 3 ??)$  shown in blue matches here with the hole  $?? = (* (+ 2 4) 2)$ , and the complete abstraction  $(+ 3 (* \alpha \beta))$  matches here with  $\alpha = (+ 2 4)$  and  $\beta = 2$ . For complete abstractions, matching corresponds to being able to use the abstraction to rewrite this subtree, resulting in compression. We refer to the set of subtrees at which a complete or partial abstraction matches as the set of *match locations*. For example, the three match locations of  $(+ 3 (* \alpha \beta))$  are the subtree  $(+ 3 (* (+ 2 4) 2))$  in the first program,  $(+ 3 (* 4 (+ 3 x)))$  in the second program, and  $(+ 3 (* x (+ 2 1)))$  in the third program.

In traditional top-down synthesis, a branch of search can be safely pruned by proving that a program satisfying the specification cannot exist in that branch. In CTS, we can safely prune a branch of search if we can prove that it cannot contain the optimal abstraction. One way to prove this is by computing an upper bound on the utility of abstractions in the branch, and discarding the branch if we have previously found an abstraction with a utility that is higher than this bound. An efficient and conservative way to compute this upper bound is to overapproximate the set of match locations, then upper bound the compressive utility gain from each match location.

Our key observation to compute this bound is that during search, expanding a hole in a partial abstraction yields an abstraction that is more precise and thus matches at a *subset* of the locations that the original matched at. This *subset observation* is depicted in Fig. 2 where the refinement of the partial abstraction  $(+ 3 ??)$  (blue) to the goal  $(+ 3 (* \alpha \beta))$  (green) results in a larger abstraction that matches at a subset of the locations (match locations shown at the top of Fig. 2B). An important consequence of this is that the match locations of a partial abstraction serves as an over-approximation of the match locations of any abstraction in this branch of top-down search.

To upper bound the compressive utility gained by rewriting at a match location, notice that the most compressive abstraction that matches at a subtree is the constant abstraction corresponding to the subtree itself. For example, the largest possible abstraction at  $(+ 3 (* (+ 2 4) 2))$  is just  $(+ 3 (* (+ 2 4) 2))$ . In other words, the size of the subtree is a strict upper bound on how much compression can be achieved at a match location. Thus we get the upper bound for some partial abstraction  $A_{??}$ :

$$U_{\text{upperbound}}(A_{??}) = \sum_{e \in \text{matches}(A_{??})} \text{size}(e) \quad (4)$$

Equipped with this bound, our algorithm performs a branch-and-bound style top-down search, where at each step of search it discards all partial abstractions that have utility upper bounds that are less than the utility of the best complete abstraction found so far.

Our full algorithm presented in Section 3 has some additional complexity. It handles rewriting in the presence of variables soundly, it uses an exact utility function that accounts for additional compression gained by using the same variable in multiple places, and it handles situations where match locations overlap and preclude one another. We also introduce two other important forms of pruning while maintaining the optimality of the algorithm, and we use the upper bound as a heuristic to prioritize more promising branches of search first.

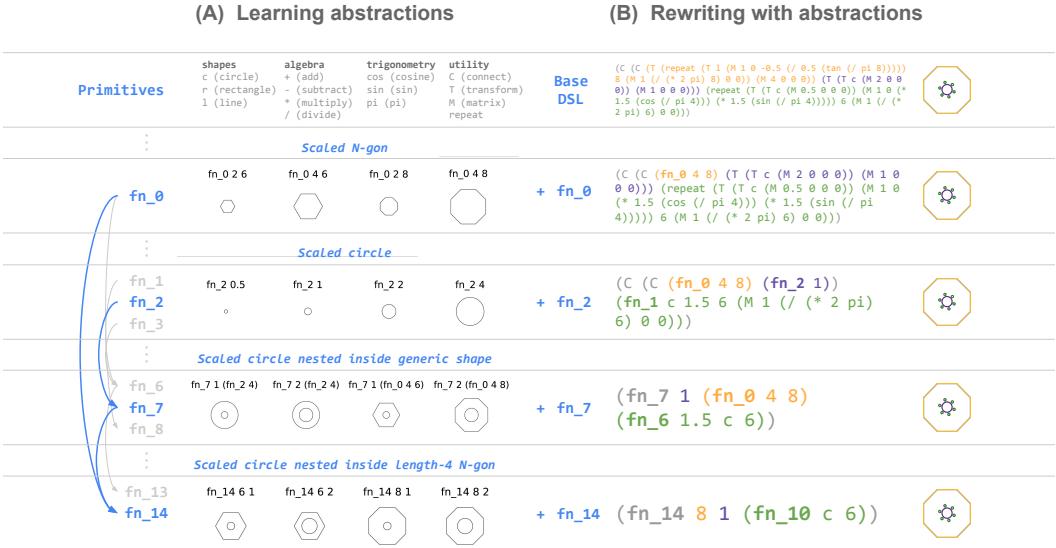


Fig. 3. Visualization of STITCH library learning on the nuts-bolts subdomain from Wong et al. 2022. (A) From the base DSL primitives (top row), STITCH iteratively discovers a series of abstractions that compress programs in the domain. Arrows demonstrate how abstractions from selected iterations build on one another to achieve increasingly higher-level behaviors. (B) Rewriting a single item from the domain with the cumulative benefit of discovered abstractions yields increasingly compact expressions. Colors indicate correspondence between object parts and program fragments: orange = outer octagon, green = ring of six circles, purple = inner circle.

**Building Up Abstraction Libraries.** The top-down search algorithm described above yields a single abstraction. However, we can easily run this algorithm for multiple iterations on a corpus of programs to build up an entire *library of abstractions*. Fig. 3 illustrates the power of this kind of iterative library learning, which interleaves compression and rewriting. At each iteration, STITCH discovers a single abstraction that is used to rewrite the entire corpus of programs. Successive iterations therefore yield abstractions that build hierarchically on one another, achieving increasingly higher-level behaviors. As the library grows to contain richer and more complex abstractions, individual programs shrink into compact expressions.

**Structure of the Paper.** In the subsequent sections, we formalize our corpus-guided top-down search algorithm (Section 3), its application to the problem of compression (Section 4), and how it may be layered on top of data structures such as version spaces (Section 5). We then report experimental results (Section 6) showcasing both diverse library learning settings as well as ablations of STITCH’s search mechanisms. Finally, we conclude by situating STITCH within the landscape of related work (Section 7) and future work (Section 8) in the areas of library learning and program synthesis.

### 3 CORPUS-GUIDED TOP-DOWN SEARCH

We first provide the definitions necessary to understand our problem then introduce our algorithm.

**Grammar.** Our algorithm operates on lambda-calculus expressions with variables represented through de Bruijn indices [de Bruijn 1972]; expressions come from a context-free grammar of the form  $e ::= \lambda. e' | e' e'' | \$i | t$ , where  $t \in \mathcal{G}_{\text{sym}}$  refers to the set of built-in primitives in the

domain-specific language. For example, in an arithmetic domain  $\mathcal{G}_{\text{sym}}$  would include operators like  $+$  and constants like 3. We say that  $e'$  is a subexpression of  $e$  if  $e = C[e']$ , where  $C$  is a context as defined in the contextual semantics of [Felleisen and Hieb 1992](#). An expression is *closed* if it has no free variables, in which case the expression is also a *program*. A set of programs  $\mathcal{P}$  is a *corpus*.

When representing variables through de Bruijn indices,  $\$i$  refers to the variable bound by the  $i$ th closest lambda above it. Thus  $\lambda x. \lambda y. (x y)$  is represented as  $\lambda. \lambda. (\$1 \$0)$ . Beta reduction with de Bruijn indices requires *upshifting*: incrementing free variables in the argument when substitution recurses into a lambda, since the free variables must now point past one more lambda. Similarly, inverse beta reduction requires *downshifting*.

**Abstraction.** Given a grammar  $\mathcal{G}$ , we define the *abstraction grammar*  $\mathcal{G}_A$  as  $\mathcal{G}$  extended to include *abstraction variables*, denoted by Greek letters  $\alpha, \beta$ , etc. Formally,  $A ::= \lambda. A' | A' A'' | \$i | t | \alpha$  where  $t \in \mathcal{G}_{\text{sym}}$ . A term  $A$  from this grammar represents the *body* of an abstraction; e.g. the abstraction  $\lambda\alpha. \lambda\beta. (+ \alpha \beta)$  is simply represented by the term  $(+ \alpha \beta)$  from the language of  $\mathcal{G}_A$ .

**Partial Abstraction.** A *partial abstraction*  $A_{??}$  is an abstraction that can additionally include *holes*, denoted by  $??$ , represents an unfinished subtree of the abstraction. Thus,  $A_{??} ::= A | ?? | \lambda. A'_{??} | A'_? A''_?$ . Any abstraction is thus also a partial abstraction. Given a grammar  $\mathcal{G}$ , the grammar of partial abstractions is denoted by  $\mathcal{G}_{A??}$ . Each hole in a partial abstraction  $A$  can be referred to by a *unique* index, which can be explicitly written as  $??_i$ .

**Lambda-Aware Unification.** We introduce *lambda-aware unification* as a modification of traditional unification adapted to our algorithm. LAMBDAUNIFY( $A, e$ ) returns a mapping (if one exists) from abstraction variables and holes to expressions  $[\alpha_i \rightarrow e'_i, \dots, ??_j \rightarrow u'_j, \dots]$  such that

$$(\lambda\alpha_i. \dots \lambda??_j. \dots A) e'_i \dots u'_j \dots = e \quad (5)$$

through beta reduction. A key difference from traditional unification is that the expression  $e$  may be deep inside a program written using de Bruijn indices, so LAMBDAUNIFY must perform some index arithmetic in order to generate its output mapping; in particular, raising a subtree out of a lambda during this inverse beta reduction requires downshifting variables in it.

The definition of LAMBDAUNIFY is presented in Fig. 4 (left). In U-APP, MERGE( $l_1, l_2$ ) merges two  $[\alpha_i \rightarrow e_i]$  mappings to create a new mapping that includes all bindings from  $l_1$  and  $l_2$ , and fails if the same abstraction variable  $\alpha$  maps to different expressions in  $l_1$  and  $l_2$ . The rule U-SAME applies only when the abstraction argument to LAMBDAUNIFY is an expression (i.e. it contains no holes or abstraction variables) and is syntactically identical to the expression it is being unified with. In U-LAM, DOWNSHIFTALL returns a new mapping with all abstracted expressions  $e$  and hole expressions  $u$  downshifted using the  $\downarrow$  operator presented in Fig. 4 (right). The  $\downarrow$  operator has been modified from traditional downshifting to allow for partial abstractions that contain holes  $??_j$  to break the rules of variable binding, because holes represent unfinished subtrees of the abstraction.

In particular, a new syntactic form " $\&i$ " is used to represent a  $\$i$  variable that has been downshifted further than traditional downshifting would permit.  $\&i$  variables are created by  $\downarrow$  when a traditional downshift would otherwise convert a free variable to a (different, incorrect) bound variable. These variables represent references to lambdas present within the body of the abstraction and are allowed in expressions  $u'_j$  bound to holes but not expressions  $e'_i$  bound to abstraction variables.

To account for  $\&i$  variables, the beta reduction used in Eq. (5), which is defined in Fig. 5, uses a modified upshift operator  $\uparrow$  defined to be an inverse to the  $\downarrow$  operator. With this modification of beta reduction, any  $\&i$  variables with negative indices will ultimately be shifted back to positive indices during reduction. Furthermore, the beta reduction procedure is equivalent to traditional beta reduction when there are no holes and thus no  $\&i$  variables.

$$\begin{array}{c}
\text{U-ABSVAR} \frac{}{\text{LAMBDAUNIFY}(\alpha, e) \rightsquigarrow [\alpha \rightarrow e]} \\
\text{U-HOLE} \frac{}{\text{LAMBDAUNIFY}(??_i, e) \rightsquigarrow [??_i \rightarrow e]} \\
\text{U-APP} \frac{\begin{array}{c} \text{LAMBDAUNIFY}(A_1, e_1) \rightsquigarrow l_1 \\ \text{LAMBDAUNIFY}(A_2, e_2) \rightsquigarrow l_2 \\ l = \text{MERGE}(l_1, l_2) \end{array}}{\text{LAMBDAUNIFY}((A_1 A_2), (e_1 e_2)) \rightsquigarrow l} \\
\text{U-LAM} \frac{\begin{array}{c} \text{LAMBDAUNIFY}(A, e) \rightsquigarrow l' \\ l = \text{DOWNSHIFTALL}(l') \end{array}}{\text{LAMBDAUNIFY}((\lambda. A), (\lambda. e)) \rightsquigarrow l} \\
\text{U-SAME} \frac{}{\text{LAMBDAUNIFY}(e, e) \rightsquigarrow []}
\end{array}
\qquad
\begin{array}{l}
\text{DOWNSHIFTALL}([\alpha_i \rightarrow e'_i, ??_j \rightarrow u'_j, \dots]) \\
= [\alpha_i \rightarrow \downarrow_0 e'_i, ??_j \rightarrow \downarrow_0 u'_j, \dots] \\
\downarrow_d \lambda.b = \lambda. \downarrow_{d+1} b \\
\downarrow_d (fx) = (\downarrow_d f \downarrow_d x) \\
\downarrow_d \$i = \begin{cases} \$i, & \text{if } i < d \\ \$i-1, & \text{if } i > d \\ \&(i-1), & \text{if } i = d \end{cases} \\
\downarrow_d \&i = \&(i-1), \\
\downarrow_d t = t, \text{ if } t \text{ is a primitive (i.e. } t \in \mathcal{G}_{\text{sym}}\text{)}
\end{array}$$

Fig. 4. (Left) Inference rules for lambda-aware unification. (Right) Definition of DOWNSHIFTALL.

$$\begin{array}{c}
(\lambda \alpha_i. \dots \lambda ??_j. \dots A) e'_i \dots u'_j \dots \\
= [\alpha_i \rightarrow e'_i, ??_j \rightarrow u'_j, \dots] \circ A \\
l \circ \lambda.b = \lambda. \text{UPSHIFTALL}(l) \circ b \\
l \circ (fx) = (l \circ f)(l \circ x) \\
l \circ \$i = \$i \\
l \circ \alpha_i = l[\alpha_i] \\
l \circ ??_i = l[??_i] \\
l \circ t = t, \text{ for } t \in \mathcal{G}_{\text{sym}}
\end{array}
\qquad
\begin{array}{l}
\text{UPSHIFTALL}([\alpha_i \rightarrow e'_i, ??_j \rightarrow u'_j, \dots]) \\
= [\alpha_i \rightarrow \uparrow_0 e'_i, ??_j \rightarrow \uparrow_0 u'_j, \dots] \\
\uparrow_d \lambda.b = \lambda. \uparrow_{d+1} b \\
\uparrow_d (fx) = (\uparrow_d f \uparrow_d x) \\
\uparrow_d \$i = \begin{cases} \$i, & \text{if } i < d \\ \$i+1, & \text{if } i \geq d \end{cases} \\
\uparrow \&i = \begin{cases} \&(i+1), & \text{if } i+1 \neq d \\ \$i+1, & \text{if } i+1 = d \end{cases} \\
\uparrow_d t = t, \text{ for } t \in \mathcal{G}_{\text{sym}}
\end{array}$$

Fig. 5. (Left) Definition of modified beta reduction and substitution ( $\circ$ ). (Right) Definition of UPSHIFTALL.

We provide a proof of the correctness of LAMBDAUNIFY with respect to Eq. (5) in Appendix B, as well as a Coq proof of the correctness in `stitch.v` in the supplemental material. However, to understand the key idea behind the proof we focus here on an example illustrating the different cases involving the  $\downarrow$  operator and  $\&i$  indices. Consider what happens when you run

$$\text{LAMBDAUNIFY}(\lambda. f ??_0, \lambda. e) \tag{6}$$

By Eq. (5), the goal is to produce a mapping  $[??_0 \rightarrow u'_0]$  such than when replacing  $??_0$  with  $u'_0$  it produces  $(\lambda. e)$ . In other words,  $(\lambda. \lambda. f \$1) u'_0 = \lambda. e$ . Now, suppose

$$\text{LAMBDAUNIFY}(f ??_0, e) \rightsquigarrow [??_0 \rightarrow \lambda. \$i] \tag{7}$$

This means that  $(\lambda. f \$0) (\lambda. \$i) = e$  so  $e = f (\lambda. \$i)$ . There are three possibilities for  $i$  that need to be considered, corresponding to the 3 cases in the definition of  $\downarrow_d \$i$ . In the first case, when  $i = 0$  and thus  $e = f (\lambda. \$0)$ , it means that  $\$i$  in Eq. (7) is bound to the lambda in the return mapping. In this case, DOWNSHIFTALL in U-LAM should not do anything because if we can replace  $??_0$  with  $(\lambda. \$0)$  in  $(f ??_0)$  to produce  $e$ , the same replacement in  $(\lambda. f ??_0)$  will produce  $(\lambda. e)$ .

The second case is when  $i \geq 2$ ; this means that  $e$  has some variables that were defined outside of it and they have been captured by  $\text{??}_0$ . For example, if  $i = 2$  and thus  $e = f(\lambda. \$2)$ , the solution to Eq. (6) is  $[\text{??}_0 \rightarrow \lambda. \$1]$ , since  $(\lambda. \lambda. f \$1)(\lambda. \$1) = \lambda. e$ ; in other words, when computing  $\text{LAMBDAUNIFY}(\lambda. f \text{??}_0, \lambda. e)$  from Eq. (7), the  $\$i$  in the return mapping has to be downshifted to account for the fact that it will be substituted inside an additional lambda.

The third case, when  $i = 1$ , is more problematic. In this case, we have that  $e = f(\lambda. \$1)$ . This creates a problem, since there is no  $i$  such that  $(\lambda. \lambda. f \$1)(\lambda. \$i) = \lambda. e$ ; the  $\$i$  needs to refer to the lambda surrounding  $e$ . The downshift operator addresses this through the special treatment of the  $\&i$  index.

$\text{LAMBDAUNIFY}$  relates to prior work on unification modulo binders [Dowek et al. 1995, 1996; Huet 1975; Miller 1991, 1992] but focuses on a more efficiently solvable syntax-driven subset of the more general unification problems tackled in this prior work. We give a more detailed comparison to this prior work in Section 7.

**Match Locations.** The set of *match locations* of a partial abstraction  $A_{??}$  in a corpus  $\mathcal{P}$ , denoted  $\text{MATCHES}(\mathcal{P}, A_{??})$ , is the maximal set of context-expression pairs  $\{(C_1, e_1), (C_2, e_2), \dots, (C_k, e_k)\}$  such that  $\forall k. p = C_k[e_k]$  and  $p \in \mathcal{P}$  and  $\text{LAMBDAUNIFY}(e_k, A_{??})$  succeeds. We discard locations where the mapping produced by  $\text{LAMBDAUNIFY}$  has  $\&i$  indices in expressions bound by abstraction variables, while we allow them in expressions bound by holes as holes are allowed to violate variable binding rules. In Section 3.1 we explain how we maintain the set of matches incrementally.

**Rewrite Strategies.** A corpus  $\mathcal{P}$  can be *rewritten* to use a complete abstraction  $A$  as follows. We introduce a new terminal symbol  $t_A$  into the symbol grammar  $\mathcal{G}_{\text{sym}}$  to represent the abstraction, and consider it semantically equivalent to  $(\lambda \alpha_0. \dots \lambda \alpha_k. A)$ . We then re-express  $\mathcal{P}$  in terms of  $t_A$  as follows. We can replace the match location  $(C, e) \in \text{MATCHES}(\mathcal{P}, A)$  with  $t_A$  applied with the argument assignments  $[\alpha_i \rightarrow e_i] = \text{LAMBDAUNIFY}(A, e)$  to yield a semantically equivalent expression,  $C[(t_A e_0 \dots e_k)]$ . Note that since this is a complete abstraction, it contains no holes.

One complication is that rewriting at one match location may *preclude* rewriting at another match location if they overlap with each other. A *rewrite strategy*  $\mathcal{R}$  is a procedure for selecting a subset of  $\text{MATCHES}(\mathcal{P}, A)$  to rewrite at, in which no match precludes another match. We refer to this set as  $\text{REWRITELocations}_{\mathcal{R}}(\mathcal{P}, A)$ . We refer to rewriting a corpus  $\mathcal{P}$  under an abstraction  $A$  with rewrite strategy  $\mathcal{R}$  by  $\text{REWRITE}_{\mathcal{R}}(\mathcal{P}, A)$ . We refer to rewriting at only a single particular match location  $m$  with abstraction  $A$  by  $\text{REWRITEONE}(m, A)$ .

**Utility.** CTS optimizes a user-defined utility function  $U_{\mathcal{P}, \mathcal{R}}(A)$  which scores an abstraction  $A$  given a corpus  $\mathcal{P}$  and rewrite strategy  $\mathcal{R}$ . There are no strict constraints on the form or properties of the utility function. While in Section 4 we will focus on compressive utility functions, our framework does not generally require this.

We say that a rewrite strategy is *optimal* with respect to a utility function if the rewrite strategy chooses locations to rewrite at such that the utility is maximized. A naïve optimal rewrite strategy exhaustively checks the utility of rewriting with each of the  $2^{|\text{MATCHES}(\mathcal{P}, A)|}$  possible subsets of  $\text{MATCHES}(\mathcal{P}, A)$ , however depending on the specific utility function computing an optimal or approximately optimal strategy may be significantly more computationally tractable. In Section 4 we detail the rewrite strategy used by STITCH, which is an optimal, linear-time rewrite strategy for compressive utility functions based on dynamic programming.

### 3.1 Algorithm

Given a corpus  $\mathcal{P}$ , rewrite strategy  $\mathcal{R}$ , and utility function  $U_{\mathcal{P}, \mathcal{R}}(A)$ , the objective of CTS is to find the abstraction  $A$  that maximizes the utility  $U_{\mathcal{P}, \mathcal{R}}(A)$ . CTS takes a branch-and-bound approach [Land and Doig 1960; Morrison et al. 2016] to this problem, as described in this section.

**Expansion.** We construct the body of a partial abstraction  $A_{??}$  through a series of top-down *expansions* starting from the trivial partial abstraction  $??$ . Given a partial abstraction  $A_{??}$  we can *expand* a hole in this partial abstraction using any production rule from the partial abstraction grammar  $\mathcal{G}_{A_{??}}$ , yielding a new abstraction  $A'_{??}$ . We denote this expansion by  $A_{??} \rightarrow A'_{??}$ . In Figure 2A our overall top-down search is depicted as a series of expansions. We say that a complete abstraction  $A$  can be *derived* from a partial abstraction  $A_{??}$ , denoted  $A_{??} \rightarrow^* A$ , if  $A$  is in the transitive reflexive closure of the expansion operation, i.e. there exists a sequence of expansions from  $A_{??}$  to  $A$ . Any abstraction can be derived from  $??$ .

**A Naïve Approach.** The goal of our algorithm is to find the maximum-utility abstraction. One simple, inefficient approach to this is to simply enumerate the *entire* space of abstractions through top down synthesis and return the one with the highest utility. This naïve approach maintains a queue of partial abstractions initialized with just  $??$ . At each step of the algorithm it pops an abstraction from the queue, chooses a hole in it and expands that hole using each possible production rule. Whenever an expansion produces a new partial abstraction it pushes it to the queue, and when it produces a complete abstraction it calculates its utility and updates the best abstraction found so far. Since an abstraction cannot match a program that is smaller than it, the algorithm can stop expanding when all expansions would lead to abstractions larger than the largest program in the corpus. This algorithm will enumerate all possible abstractions exactly once each.

**Introducing Strict Dominance Pruning.** While the naïve approach will find the optimal abstraction, it is extremely inefficient. We can improve on this using an idea core to branch-and-bound algorithms: *pruning*. We allow the algorithm to choose to prune a partial abstraction instead of expanding it, in which case it simply discards the abstraction and chooses another from the queue. Of course, to maintain optimality we must be certain that we *never* prune the branch of search containing the optimal abstraction. To formalize safe pruning we will use the notion of *strict dominance* from branch-and-bound literature [Chu and Stuckey 2015; Ibaraki 1977; Morrison et al. 2016], which we explain below in terms of a notion of *covering*.

A complete abstraction  $A''$  *covers* another complete abstraction  $A'$ , written  $\text{covers}(A'', A')$ , if the utility of  $A''$  is greater than that of  $A'$ . A partial abstraction  $A'_{??}$  is *strictly dominated* by another partial abstraction  $A''_{??}$  if and only if every complete abstraction that is derivable from  $A'_{??}$  is covered by a complete abstraction that is derivable from  $A''_{??}$ .

Formally,  $A''_{??}$  strictly dominates  $A'_{??}$  iff  $\forall A'. A'_{??} \rightarrow^* A' \implies \exists A''. A''_{??} \rightarrow^* A'' \wedge \text{covers}(A'', A')$ .

Equipped with this formalism, we claim that it is safe to prune an abstraction  $A'_{??}$  if we know that there exists some  $A''_{??}$  which strictly dominates it. Note that  $A''_{??}$  does not necessarily contain the optimum and may even be pruned in the search if it is strictly dominated by another abstraction. Knowing the existence of  $A''_{??}$  is enough to enable pruning, regardless of whether it has been pruned or has not yet been enumerated during search.

**LEMMA 1.** *Naïve search augmented with strict dominance pruning finds the optimal abstraction.*

**PROOF.** We proceed with a proof by induction. We seek to prove the equivalent statement that the optimum is never pruned and thus it will be enumerated by the search.

Our inductive hypothesis is that the optimum has not yet been pruned. In the base case this is trivially true, since no pruning has taken place yet. In the inductive step we must prove that a

step of pruning maintains the validity of the inductive hypothesis. Recall that a step of pruning will discard some partial abstraction  $A'_{??}$  which is strictly dominated by another abstraction  $A''_{??}$ . Suppose then for the sake of contradiction that we did prune the optimum in this step; then the optimum must have been derivable from  $A'_{??}$ . By strict dominance we know that all abstractions derivable from  $A'_{??}$ , must be covered by some abstraction derivable from  $A''_{??}$ . Hence, there must exist some abstraction that covers the optimum. However, since the utility of the optimum is greater than or equal to that of all other abstractions, there is by definition *no* such abstraction; we have thus arrived at a contradiction.  $\square$

This lemma ensures the safety of composing together pruning strategies, as long as they all are forms of strict dominance pruning, since no instance of strict dominance pruning will remove the optimum. Determining which partial abstractions are strictly dominated by which others is specific to the utility function being used, and in Section 4 we identify two instances of strict dominance used when instantiating this framework for a compression-based utility.

**Upper Bound Pruning.** We further improve this algorithm to employ upper bound based pruning, in which we bound the maximum utility that can be obtained in a branch of search, and discard the branch if we have previously found a complete abstraction with higher utility than this bound. This is the most common form of pruning used in branch-and-bound algorithms (see Morrison et al. 2016 Section 5.1 for a review). Our algorithm is generic over the upper bound function  $\bar{U}_{\mathcal{P}, \mathcal{R}}(A_{??})$  which upper bounds the utility of *any complete abstraction A* that can be derived from  $A_{??}$ .

Formally, the bound must satisfy  $\forall A. A_{??} \rightarrow^* A \implies U_{\mathcal{P}, \mathcal{R}}(A) \leq \bar{U}_{\mathcal{P}, \mathcal{R}}(A_{??})$ . While this could trivially be satisfied with  $\bar{U}_{\mathcal{P}, \mathcal{R}}(A_{??}) = \infty$  for any choice of a utility function, a tighter bound will allow for more pruning. The soundness of upper bound pruning in maintaining the optimality of the solution trivially follows from the fact that the pruned branch only contains complete abstractions that are at most as high-utility as the best abstraction found so far.

While we defer to Section 4 to construct the upper bound used by STITCH, it is worth mentioning a key insight that helps in constructing tight bounds for many utility functions. When bounding the utility of  $A_{??}$ , it is useful to bound the set of possible locations where rewrites can occur for any abstraction  $A$  derived from  $A_{??}$ . The following lemma provides such a bound:

**LEMMA 2.** *MATCHES( $\mathcal{P}, A_{??}$ ) is an upper bound on the set of locations where rewrites can occur in any A derived from  $A_{??}$ .*

This follows from the fact that as partial abstractions are expanded they become more precise and thus match at a subset of the locations. Since rewriting only happens at a subset of match locations, MATCHES( $\mathcal{P}, A_{??}$ ) serves as a bound on the set of locations where rewrites can occur.

An important consequence of Lemma 2 is that if a partial abstraction matches at zero locations, then all abstractions derived from it will have zero rewriting locations, so no rewriting can occur. Such branches can therefore be safely pruned.

**Improving the Search Order.** Finally, without sacrificing the optimality of this algorithm we can heuristically guide the order in which the space of abstractions is explored. The queue of partial abstractions used in top-down search can be replaced with a priority queue ordered by the upper bound. This way, the algorithm will first explore more promising, higher-bound branches of search, narrowing in on the optimal abstraction more quickly. In branch-and-bound literature this choice of a search order that uses a sound upper bound is sometimes referred to as *best-bound* search and makes the algorithm a form of  $A^*$  search [Hart et al. 1968] since the upper bound is an admissible heuristic [Morrison et al. 2016].

In preliminary experiments on a subset of benchmarks we found that search order had little effect on the overall runtime of the algorithm, but the best-bound ordering was moderately helpful in more quickly narrowing in on the optimal solution (i.e. useful when running the algorithm with a limited time budget), so this is used for all experiments.

**Efficient Incremental Matching.** When expanding an abstraction  $A_{??}$  to a new abstraction  $A'_{??}$ , it is easy to compute  $\text{MATCHES}(\mathcal{P}, A'_{??})$  since we know it is a subset of  $\text{MATCHES}(\mathcal{P}, A_{??})$ . Thus, there is no need to perform matching from scratch against every subtree in the corpus to compute the match locations. Instead, we can simply inspect the relevant subtree at each match location of the original abstraction to see which match locations will be preserved by a given expansion. In fact, except when expanding into an abstraction variable  $\alpha$ , the sets of match locations obtained by different expansions of a hole will be disjoint.

When expanding to an abstraction variable  $\alpha$ , if  $\alpha$  is an existing abstraction variable then this is a situation where the same variable is being used in more than one place, as in the *square* abstraction  $(\lambda \alpha. * \alpha \alpha)$ . In this case we restrict the match locations to the subset of locations where within the location all instances of  $\alpha$  are bound to syntactically identical subtrees.

Additionally, if the user provides a maximum arity limit, then an expansion that causes the abstraction to exceed this limit can be discarded.

**Algorithm Summary and Key Points.** The CTS algorithm combines all of the aforementioned optimizations in a top-down search algorithm with pruning. In summary, CTS takes as input a corpus  $\mathcal{P}$ , a rewrite strategy  $\mathcal{R}$ , a utility function  $U_{\mathcal{P}, \mathcal{R}}(A)$ , and an upper bound function  $\bar{U}_{\mathcal{P}, \mathcal{R}}(A_{??})$ , and returns the *optimal abstraction* with respect to the utility function. CTS performs a top-down search over partial abstractions and prunes branches of search when partial abstractions match at zero locations or are eliminated by upper bound pruning or strict dominance pruning. CTS is made further efficient through the aforementioned *search-order heuristic* and *efficient incremental matching*. Importantly, none of these optimizations sacrifice the optimality of the abstraction found – all pruning is done soundly, as discussed in the prior sections on upper bound pruning and strict dominance pruning.

We also note that CTS is amenable to parallelization without losing optimality. This can be implemented using a shared priority queue that is safely accessed by different worker threads through a locking primitive. Strict dominance pruning trivially remains sound as it doesn't depend on any global information about the state of the search. Upper bound pruning also remains sound even if workers only occasionally synchronize their best-abstraction-so-far, as this will just mean that they occasionally have weaker upper bounds which are still sound.

Since the algorithm maintains a best abstraction so-far, it can also be terminated early, making it an *anytime algorithm*. Finally, to learn a library of abstractions, CTS can be run repeatedly (much like DreamCoder), adding one abstraction at a time to the library and rewriting the corpus with each abstraction as it is learned before running CTS again. A listing of the full algorithm is provided in Appendix A.

## 4 APPLYING CORPUS-GUIDED TOP-DOWN SEARCH TO COMPRESSION

Having presented the general framework and algorithm of corpus-guided top-down search, we now instantiate this framework for optimizing a *compression* metric.

### 4.1 Utility

In compression we seek to minimize some measure of the size, or more generally the *cost*, of a corpus of programs after rewriting them with a new abstraction. As in prior work [Ellis et al. 2021]

we penalize large abstractions by including abstraction size in the utility. The compressive utility function for a corpus  $\mathcal{P}$  and abstraction  $A$  is given below.

$$U_{\mathcal{P}, \mathcal{R}}(A) \triangleq -\text{cost}(A) + \text{cost}(\mathcal{P}) - \text{cost}(\text{REWRITE}_{\mathcal{R}}(\mathcal{P}, A)) \quad (8)$$

Here,  $\text{cost}(\cdot)$  is a cost function of the following form:

$$\begin{aligned} \text{cost}(\lambda. e') &= \text{cost}_{\lambda} + \text{cost}(e') \\ \text{cost}(e'e'') &= \text{cost}_{\text{app}} + \text{cost}(e') + \text{cost}(e'') \\ \text{cost}(\$i) &= \text{cost}_{\$i} \\ \text{cost}(t) &= \text{cost}_t(t), \text{ for } t \in \mathcal{G}_{\text{sym}} \\ \text{cost}(\alpha) &= \text{cost}_{\alpha} \end{aligned} \quad (9)$$

where  $\text{cost}_{\lambda}$ ,  $\text{cost}_{\text{app}}$ ,  $\text{cost}_{\$i}$ , and  $\text{cost}_{\alpha}$  are non-negative constants, and  $\text{cost}_t(t)$  is a mapping from grammar primitives to their (non-negative) costs. Finally, we introduce  $\text{cost}_{\alpha=0}(\cdot)$  as a version of  $\text{cost}(\cdot)$  where  $\text{cost}_{\alpha} = 0$ .

We can equivalently construct the utility in Eq. (8) by summing over the compression gained from performing each rewrite separately, as given below.

$$U_{\mathcal{P}, \mathcal{R}}(A) = -\text{cost}(A) + \sum_{e \in \text{REWRITELOCATIONS}_{\mathcal{R}}(\mathcal{P}, A)} \text{cost}(e) - \text{cost}(\text{REWRITEONE}(e, A)) \quad (10)$$

By reasoning about the way that rewriting transforms a program, this utility can be broken down even further:

$$\begin{aligned} U_{\mathcal{P}, \mathcal{R}}(A) &= -\text{cost}(A) + \sum_{e \in \text{REWRITELOCATIONS}_{\mathcal{R}}(\mathcal{P}, A)} U_{\text{local}}(A, e) \\ U_{\text{local}}(A, e) &= \underbrace{\text{cost}_{\alpha=0}(A) - (\text{cost}_t(t_A) + \text{cost}_{\text{app}} \cdot \text{arity}(A))}_{\text{abstraction size}} + \underbrace{\sum_{[\alpha \rightarrow e'] \in \text{args}(A, e)} (\text{usages}(\alpha) - 1) \cdot \text{cost}(e')}_{\text{application utility (negative)}} + \underbrace{\sum_{[\alpha \rightarrow e'] \in \text{args}(A, e)} (\text{usages}(\alpha) - 1) \cdot \text{cost}(e')}_{\text{multiuse utility}} \end{aligned} \quad (11)$$

where  $t_A$  is the new primitive corresponding to abstraction  $A$ . This form of the utility function can be efficiently computed without explicitly performing any rewrites; and it sheds light on the different sources of compression. The three main terms in this expression are (1) the *abstraction size* that comes from the shared structure that is removed, (2) the negative *application utility* that comes from introducing the new primitive and lambda calculus app nonterminals to apply it to each argument, and (3) the *multiuse utility* which comes from only needing to pass in a single copy of an argument that might be used in more than one place in the body. We emphasize that this form of the utility function is equivalent to the original definition based on rewriting given in Eq. (8).

## 4.2 Upper Bounding the Utility

We seek an upper bound function  $\bar{U}_{\mathcal{P}, \mathcal{R}}(A??)$  such that for any  $A$  derived from  $A??$ ,  $\bar{U}_{\mathcal{P}, \mathcal{R}}(A??) \geq U_{\mathcal{P}, \mathcal{R}}(A)$ . We begin from the decomposition of the utility function given in Eq. (10). Since costs are always non-negative, we can upper bound this by dropping the  $-\text{cost}(A)$  term as well as the negative term within the sum:

$$U_{\mathcal{P}, \mathcal{R}}(A) \leq \sum_{e \in \text{REWRITELOCATIONS}_{\mathcal{R}}(\mathcal{P}, A)} \text{cost}(e) \quad (13)$$

Intuitively, dropping the negative term from the sum is equivalent to assuming we compressed the cost of this location all the way down to cost 0. We can also bound  $\text{REWRITELocations}_{\mathcal{R}}(\mathcal{P}, A)$  as  $\text{MATCHES}(\mathcal{P}, A_{??})$  using Lemma 2, yielding our final upper bound in terms of  $A_{??}$ :

$$\bar{U}_{\mathcal{P}, \mathcal{R}}(A_{??}) \triangleq \sum_{e \in \text{MATCHES}(\mathcal{P}, A_{??})} \text{cost}(e) \quad (14)$$

### 4.3 Strict Dominance Pruning

We identify two forms of strict dominance pruning compatible with a compressive utility. The first is *redundant argument elimination*: a partial abstraction can be dropped if it has two abstraction variables that always take the same argument as each other across all match locations, for example if it had variables  $\alpha$  and  $\beta$  and  $\alpha = \beta = (+ 3 5)$  at one location,  $\alpha = \beta = 2$  at another location, and so on for all match locations. This abstraction would then be strictly dominated by the abstraction that doesn't take  $\beta$  as an argument and instead reuses  $\alpha$  in place of  $\beta$ , and can hence be eliminated. Thus an abstraction like  $(+ (* \alpha \beta) ??)$  is strictly dominated by an abstraction like  $(+ (* \alpha \alpha) ??)$  if  $\alpha = \beta$  across all match locations in the former abstraction.

The second is *argument capture*: when a partial abstraction takes the same argument for at least one abstraction variable across all match locations, this abstraction can be discarded. This is because every abstraction derivable is covered by another abstraction which is identical except it has the argument inlined into the body. For example, if in the abstraction  $(+ (* \alpha \beta) ??)$  the abstraction variable  $\alpha$  takes the same argument  $\alpha = (+ 3 5)$  across every match location, there is a strictly dominating partial abstraction that simply has  $(+ 3 5)$  inlined in its body:  $(+ (* (+ 3 5) \beta) ??)$ . Note that argument capture does not apply when the argument contains a free variable, as inlining would result in an invalid abstraction.

A close reader might notice that if the number of times  $\alpha$  appears in  $A$  is greater than the number of rewrite locations, the abstraction with argument capture applied,  $A'$ , will have slightly lower utility. Employing this pruning rule in our search means that we will find the optimal abstraction subject to the constraint that all possible argument captures have occurred. This utility difference from the optimal abstraction without argument capture is bounded by at most  $\text{cost}(A')$ , since the difference comes from the contribution of the size of the abstraction body itself to the utility.<sup>2</sup> In preliminary experiments across all of our experimental datasets, we never find this edge case to change which abstraction is chosen as optimal.

### 4.4 Rewrite Strategy

STITCH employs a linear-time, optimal rewrite strategy for a compression objective. The goal of an optimal rewrite strategy is to efficiently select the optimal subset of  $\text{MATCHES}(\mathcal{P}, A)$  to perform rewrites at in order to maximize the utility.

The main challenge is that when match locations overlap, the strategy must decide which of the two to accept. For example, consider the program  $(\text{foo} (\text{foo} (\text{foo bar}))$  and the abstraction  $t_A = \lambda \alpha. (\text{foo} (\text{foo} \alpha))$ . This abstraction matches at the root of the program with  $\alpha = (\text{foo bar})$ , resulting in the rewritten program  $(t_A (\text{foo bar}))$ . However, though the abstraction also matched at the subtree  $(\text{foo} (\text{foo bar}))$  with  $\alpha = \text{bar}$  in the original program, this match location is no longer present in the rewritten program, so only one of the two locations can be chosen by the rewrite strategy.

---

<sup>2</sup>Specifically, for some  $[\alpha \rightarrow e]$ , the abstraction without argument capture is higher in utility by  $-(\text{cost}(e) + \text{cost}_{\text{app}}) * |\text{REWRITELocations}_{\mathcal{R}}(\mathcal{P}, A)| + (\text{cost}(e) - \text{cost}_\alpha) * \text{usages}(\alpha)$ . For a cost function where  $\text{cost}_{\text{app}} < \text{cost}_\alpha$ , this is positive when the argument is used more times than the abstraction itself is used.

Our approach is a bottom-up dynamic programming algorithm, which begins at the leaves of the program and proceeds upwards. At each node  $e$ , we compute the cumulative utility so far if we reject a rewrite here ( $\text{util}_R$ ), if we accept a rewrite here ( $\text{util}_A$ ), or if we choose the better of the two options ( $\text{util}^*$ ):

$$\begin{aligned}\text{util}_R[e] &= \sum_{e' \in \text{children}(e)} \text{util}^*[e'] \\ \text{util}_A[e] &= \begin{cases} 0 & \text{if } e \notin \text{MATCHES}(\mathcal{P}, A) \\ U_{\text{local}}(A, e) + \sum_{e' \in \text{args}(A, e)} \text{util}^*[e'] & \text{otherwise} \end{cases} \\ \text{util}^*[e] &= \max(\text{util}_R[e], \text{util}_A[e])\end{aligned}\tag{15}$$

where  $U_{\text{local}}(A, e)$  is the utility gained from a single rewrite as defined in Eq. (12). After calculating these quantities the rewrite strategy can start from the program root and recurse down the tree, rewriting at each node where rewriting is optimal (i.e.  $\text{util}_A > \text{util}_R$ ) and then recursing into its arguments after rewriting. Since all arguments were originally subtrees in the program, these quantities will have been calculated for all of them, with the caveat that their de Bruijn indices may have been shifted. Shifting indices of a subtree does not change whether an abstraction can match there nor does it change the utility gained from using that abstraction, so this simply requires some extra bookkeeping to track.

This algorithm is optimal by a simple inductive argument: at each step of the dynamic programming problem, we can assume that we know the cumulative utility of all children and potential arguments at this node, so we can use Eq. (15) to calculate the cumulative utility of either rejecting or accepting the rewrite at this node.

## 5 COMBINING CORPUS-GUIDED TOP-DOWN SEARCH WITH DEDUCTIVE APPROACHES

In complex domains, assembling good libraries may involve more than just finding matching code-templates; sometimes, some initial refactoring is necessary to expose common structure. For example, consider learning the abstraction  $\lambda\alpha. (* \ 2 \ \alpha)$  for doubling integers, given the expressions  $(* \ 2 \ 8)$ ,  $(* \ 7 \ 2)$ , and  $(\text{right-shift} \ 3 \ 1)$ . This is only possible if the system can use the commutativity of multiplication to rewrite  $(* \ 7 \ 2)$  into  $(* \ 2 \ 7)$ , and bitvector properties to rewrite  $(\text{right-shift} \ 3 \ 1)$  into  $(* \ 2 \ 3)$ ; such rewrites are not natively supported by CTS.

In this section, we discuss how CTS can be combined with refactoring systems based on deductive rewrites to increase its expressivity further. The core idea is simple: run a rewrite system on the corpus to produce a set of refactorings of the corpus in a version space, then run CTS over the resulting data structure.<sup>3</sup> Intuitively, this will lead to improved performance compared with a purely deductive approach since the cost of rewriting grows exponentially with the number iterations of rewrites that are applied. This is a problem for fully deductive approaches like Ellis et al. 2021 because extracting the abstractions often requires a long chain of rewrites, especially for higher-arity abstractions. However, a small number of rewrites is typically sufficient to expose the underlying commonalities, as it was in the example above; performing only a handful of rewrites and then using CTS to actually extract the library thus avoids the exponential blow-up of computing several rewrites in sequence, while still benefiting from the increase in expressivity afforded by the rewrites.

---

<sup>3</sup>Note that while we have previously presented CTS as operating over syntax trees, the core notions of upper bounds and matching that CTS operates on are not restricted to program trees.

$$\begin{array}{c}
\text{RU-ABSVAR} \frac{}{\text{REWRITEUNIFY}(\alpha, v) \rightsquigarrow [\alpha \rightarrow v]} \quad \text{RU-HOLE} \frac{}{\text{REWRITEUNIFY}(??_i, v) \rightsquigarrow [??_i \rightarrow v]} \\
\\
\text{RU-APP} \frac{\text{REWRITEUNIFY}(A_1, v_1) \rightsquigarrow l_1 \quad \text{REWRITEUNIFY}(A_2, v_2) \rightsquigarrow l_2 \quad l = \text{VMERGE}(l_1, l_2)}{\text{REWRITEUNIFY}((A_1 A_2), (v_1 v_2)) \rightsquigarrow l} \\
\\
\text{RU-LAM} \frac{\text{REWRITEUNIFY}(A, v) \rightsquigarrow l' \quad l = \text{DOWNSHIFTALL}(l')}{\text{REWRITEUNIFY}((\lambda. A), (\lambda. v)) \rightsquigarrow l} \quad \text{RU-SAME} \frac{}{\text{REWRITEUNIFY}(v, v) \rightsquigarrow []} \\
\\
\text{RU-UNION} \frac{\text{REWRITEUNIFY}(A, v_1) \rightsquigarrow l}{\text{REWRITEUNIFY}(A, v_1 \uplus v_2) \rightsquigarrow l} \\
\\
\text{VM-SAME} \frac{[\alpha \rightarrow v_1] \in l_1 \quad [\alpha \rightarrow v_2] \in l_2}{[\alpha \rightarrow (v_1 \cap v_2)] \in \text{VMERGE}(l_1, l_2)} \quad \text{VM-DIFF} \frac{[\alpha \rightarrow v] \in l_1 \quad \alpha \notin \text{args}(l_2)}{[\alpha \rightarrow v] \in \text{VMERGE}(l_1, l_2)}
\end{array}$$

Fig. 6. Defining  $\text{REWRITEUNIFY}$ , which takes as input an abstraction and a version space of possible refactorings, and yields multiple substitutions corresponding to all the ways that the abstraction’s holes and variables can match with programs encoded by the version space.

**Version Spaces.** To illustrate this hybrid approach, we combine CTS with *version spaces* [Lau et al. 2003; Polozov and Gulwani 2015] representing sets of programs semantically equivalent under the  $\beta$ -inversion rewrite. Version spaces are represented as terms from a grammar obtained by extending the grammar of expressions with the union ( $\uplus$ ) operator that represents a set of equivalent expressions:  $v ::= \emptyset \mid v' \uplus v'' \mid \lambda. v' \mid v' v'' \mid \$i \mid t$ . We define a denotation operator,  $\llbracket v \rrbracket$ , mapping a version space  $v$  to a set of terms: for the union operator,  $\llbracket v \uplus v' \rrbracket = \llbracket v \rrbracket \cup \llbracket v' \rrbracket$ ; for applications,  $\llbracket v v' \rrbracket = \{e e' : \forall e, e' \in \llbracket v \rrbracket \times \llbracket v' \rrbracket\}$ ; for lambda abstractions,  $\llbracket \lambda v \rrbracket = \{\lambda e : \forall e \in \llbracket v \rrbracket\}$ ; for the empty set,  $\llbracket \emptyset \rrbracket = \emptyset$ ; and for de Bruijn indices and terminals,  $\llbracket v \rrbracket = v$ . Ellis et al. 2021 gives a procedure called  $I\beta(e)$ , which take as input an expression  $e$  and outputs a version space  $v$  inverting one step of  $\beta$ -reduction: that is,  $e' \rightarrow^\beta e$  iff  $e' \in \llbracket I\beta(e) \rrbracket$ .

To run CTS on top of this rewriting system requires a generalization of LAMBDAUNIFY: instead of unifying an abstraction with a term, yielding a single substitution, we unify against a *set of terms* (a version space), yielding a set of candidate substitutions. The relation  $\text{REWRITEUNIFY}$  (Fig. 6) accomplishes this, and equipped with this relation we can express the size of a subtree  $e$  after expanding it into the version space  $I\beta(e)$  and rewriting it with abstraction  $A$  as:

$$\text{cost}(\text{REWRITE}(A, e)) = \min_{l: \text{REWRITEUNIFY}(A, I\beta(e)) \rightsquigarrow l} \text{cost}_l(t_A) + \text{cost}_{\text{app}} \cdot \text{arity}(A) + \sum_{v \in \text{args}(l)} \min_{e' \in \llbracket v \rrbracket} \text{cost}(e') \quad (16)$$

We then can approximate the utility of rewriting a corpus (based on Eq. (10)) as follows

$$U_{\mathcal{P}, \mathcal{R}}(A) \approx -\text{cost}(A) + \sum_{p \in \mathcal{P}} \max_{e \in \text{subtrees}(p)} \text{cost}(e) - \text{cost}(\text{REWRITE}(A, e)) \quad (17)$$

This utility is exact when the optimal way of rewriting the corpus using  $A$  has at most one rewrite location per program. This is because it considers the utility of the single best site at which to perform the rewrite instead of considering multiple simultaneous rewrites at different locations within a single program. We can bound this approximate utility for a partial abstraction by computing the approximate utility directly while treating any version space bound to a hole in  $\text{args}(l)$  as having zero cost.

**Example: Learning map from fold.** Consider learning the higher-order function  $\text{map} = \lambda\alpha.(\text{fold } (\lambda x.\lambda\ell. (\text{cons } (\alpha x) \ell)))$  from two example programs: doubling a list of numbers, expressed as  $(\text{fold } (\lambda x.\lambda\ell. (\text{cons } (+ x x) \ell)))$ , and decrementing a list of numbers, expressed as  $(\text{fold } (\lambda x.\lambda\ell. (\text{cons } (- x 1) \ell)))$ . Matching these programs with the  $\text{map}$  function requires re-expressing them as  $(\text{fold } (\lambda x.\lambda\ell. (\text{cons } ((\lambda z. (+ z z)) x) \ell)))$  and  $(\text{fold } (\lambda x.\lambda\ell. (\text{cons } ((\lambda z. (- z 1)) x) \ell)))$ , which matches the  $\text{map}$  function with the substitutions  $\alpha = (\lambda z. (+ z z))$  and  $\alpha = (\lambda z. (- z 1))$  respectively. These two re-expressions are performed by running a single round of  $\beta$ -inversion as a deductive rewriting step.

While the majority of our experiments will focus on evaluating the approach laid out in Section 3 and Section 4, we implement and evaluate a prototype of integrating version spaces in this way in Section 6.5.

## 6 EXPERIMENTS

In this section, we evaluate corpus-guided top-down search for library learning. Specifically, our evaluation focuses on five hypotheses about the performance of STITCH:

- (1) *STITCH learns libraries of comparable quality to those found by existing deductive library learning algorithms in prior work, while requiring significantly less resources.* In Section 6.1 we run STITCH on the library learning tasks from [Ellis et al. 2021] and directly compare STITCH to DreamCoder, the deductive algorithm introduced in that work. We find that STITCH learns libraries which usually match or exceed the baseline in quality (measured via a compression metric), while improving the resource efficiency in terms of memory usage and runtime by 2 and 3-4 orders of magnitude compared to the baseline (respectively).
- (2) *STITCH scales to corpora of programs that contain more and longer programs than would be tractable with prior work.* In Section 6.2, we evaluate STITCH’s ability to learn libraries within eight graphics domains from Wong et al. 2022, which are considerably larger and more complex than have been considered in previous work. We find that STITCH on average obtains a test set compression ratio of 2.55x-11.57x in 0.19s-60.16s, with a peak memory usage of 11.21MB-714.55MB in these domains. The problems are large enough to time out with the DreamCoder baseline.
- (3) *STITCH degrades gracefully when resource-constrained.* In Section 6.3 we investigate STITCH’s performance when run as an *anytime* algorithm; i.e., one that can be terminated early for a best-effort result if a corpus is too large or there are limits on time or memory. We reuse the eight graphics domains from Wong et al. 2022 and find that with its heuristic guidance STITCH converges upon a set of high quality abstractions very early in search, doing so within 1% of the total search time in 3 out of 8 domains and within 10% in all except one.
- (4) *All the elements of STITCH matter.* In Section 6.4, we carry out an ablation study on STITCH and find the *argument capturing* and *upper bound* pruning methods are essential to its performance, while *redundant argument elimination* also proves useful in certain domains. With all optimizations disabled, we find that STITCH cannot run in  $\leq 90$  minutes and  $\leq 50$ GB of RAM on any of the domains from Wong et al. 2022.
- (5) *STITCH is complementary to deductive rewrite-based approaches to library learning.* These prior experiments show the superior runtime performance of STITCH relative to deductive rewrite systems, but deductive systems have an important advantage over STITCH: the ability to incorporate arbitrary rewrite rules to expose more commonality among different programs and in that way discover better libraries. Such deductive approaches are especially more apt at learning higher-order abstractions. In Section 6.5, we give evidence that this expressivity

gap can be reduced by running STITCH on top of a deductive rewrite system, allowing it to learn many new abstractions while still using <2% of the compute time.

For all experiments, we parameterize STITCH’s cost( $e$ ) function (as defined in Section 4.1) as follows:  $\text{cost}_{\text{app}} = \text{cost}_\lambda = 1$ ,  $\text{cost}_{\$i} = \text{cost}_\alpha = \text{cost}_t(t) = 100$ . To avoid overfitting, DreamCoder prunes the abstractions that are only useful in programs from a single task. We add this to STITCH as well, treating each program as a separate task for datasets that don’t divide programs into tasks.

We run all experiments on a machine with two AMD EPYC 7302 processors, 64 CPUs, and 256GB of RAM. We note however that STITCH itself runs exceptionally well on a more average machine. For example, on one author’s laptop (ThinkPad X1 Carbon Gen 8), the experiments from Section 6.2 can be replicated with nearly identical runtimes, and these are the most computationally intensive experiments outside of the ablation study.

## 6.1 Iterative Bootstrapped Library Learning

*Experimental setup.* Our first experiment is designed to replicate the experiments in DreamCoder, which is the state-of-the-art in deductive library learning. DreamCoder learns libraries iteratively: the system is initialized with a low-level DSL, and then alternates between synthesizing programs (via a neurally-guided enumerative search) that solve a training corpus of inductive tasks and updating the library of abstractions available to the synthesizer. Traces from the experiments carried out by Ellis et al. 2021 are publicly available<sup>4</sup> and include all of the intermediate programs that were synthesized as well as the libraries learned from those programs.

In this experiment, we take these traces and evaluate STITCH on each instance where library learning was performed, comparing the quality of the resulting library to the original one found by DreamCoder. We also re-run DreamCoder on these same benchmarks in order to evaluate its resource usage, capturing its runtime and memory usage in the same environment as STITCH.

The library learning algorithm in Ellis et al. 2021 implements a stopping criterion to determine how *many* abstractions to retain on any given set of training programs. In our comparative experiments, we run the DreamCoder baseline first, and then match the number of abstractions learned by STITCH at each iteration to those learned by the baseline under its stopping criterion so that timing comparisons are fair. We record the total time spent both performing abstraction learning and rewriting for both STITCH and DreamCoder.

We replicate experiments on five distinct domains from Ellis et al. 2021:

- **Lists:** A functional programming domain consisting of 108 total inductive tasks.
- **Text:** A string editing domain in the style of FlashFill [Gulwani et al. 2015] consisting of 128 total inductive tasks.
- **LOGO:** A graphics domain consisting of 80 total inductive tasks.
- **Towers:** A block-tower construction domain consisting of 56 total inductive tasks.
- **Physics:** A domain for learning equations corresponding to physical laws from observations of simulated data, consisting of 60 total inductive tasks.

*Assessing library quality with a compression metric.* The standard STITCH configuration optimizes a compression metric that minimizes the size of the programs after being rewritten to use the abstraction. This is a standard metric in program synthesis, since shorter programs are frequently easier to synthesize. Optimizing against this metric is equivalent to maximizing the likelihood of the rewritten programs under a uniform PCFG.

The DreamCoder synthesizer is more sophisticated than simple enumeration; it takes as input a learned typed bigram PCFG and leverages it to synthesize programs more efficiently. When

---

<sup>4</sup>[https://github.com/mlb2251/compression\\_benchmark](https://github.com/mlb2251/compression_benchmark)

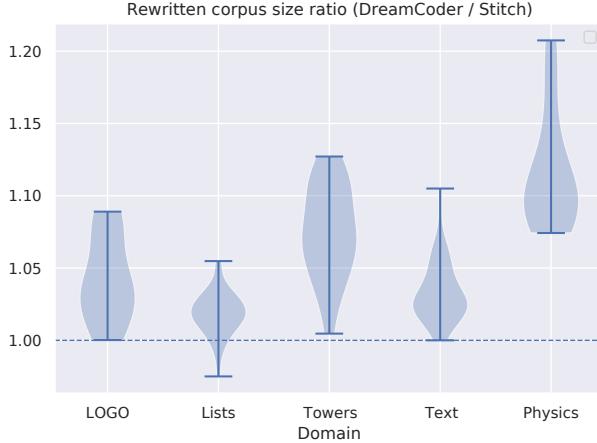


Fig. 7. Compression rates obtained when running **STITCH** on the five domains considered in 6.1 relative to those of DreamCoder. Higher is better for **STITCH**: a ratio above 1.0 indicates that **STITCH** achieves greater compression than DreamCoder. The specific compression metric used in the ratio is given in Eq. (18).

performing compression, it optimizes against this given PCFG in order to find abstractions that will be more profitable for its specific synthesizer. For the purpose of this evaluation, however, we restrict ourselves to the uniform PCFG because the one used by DreamCoder requires programs to be in a particular normal form. Another aspect of DreamCoder relevant to its compression metric is that DreamCoder synthesizes multiple programs that solve the same task and then selects the abstraction that works best on *some* program for a given task. This is expressed formally in the equation below.

$$\text{cost}(A) + \sum_{\text{task}} \min_{p \in \text{task}} \text{cost}(\text{REWRITE}(p, A)) \quad (18)$$

It is trivial to implement this best-of-task metric in **STITCH**, so we use this for the comparisons with DreamCoder.

**Results.** We compare DreamCoder and **STITCH** for library quality and resource efficiency.

**Library quality.** We first examine the quality of the libraries learned by both **STITCH** and DreamCoder using the compression metric in Eq. (18): Fig. 7 shows the ratio between them across all of the benchmarks for each domain. A ratio of 1.0 indicates that programs are the *exact same length* under the libraries learned by DreamCoder and **STITCH**, a ratio greater than 1.0 indicates that **STITCH** learns more compressive libraries, and a ratio less than 1.0 indicates that **STITCH** learns libraries which are less compressive. For example, a ratio of 1.1 indicates that DreamCoder rewrote to produce a corpus 10% larger than that of **STITCH**, so it achieved less compression.

These results show that **STITCH** generally learns libraries of comparable and often greater quality than DreamCoder when matching the number of abstractions learned by the latter. In the *logo*, *towers*, *text*, and *physics* domains, **STITCH** always finds abstractions that are of equal—and often considerably greater—compressive quality than DreamCoder does; in the *list* domain, **STITCH** more often than not still obtains better compression, but sometimes loses out to DreamCoder. This is a result of the fact that **STITCH** cannot learn higher-order abstractions, which are useful in this domain; although we emphasize our focus is on scalability, we will later present an extension of

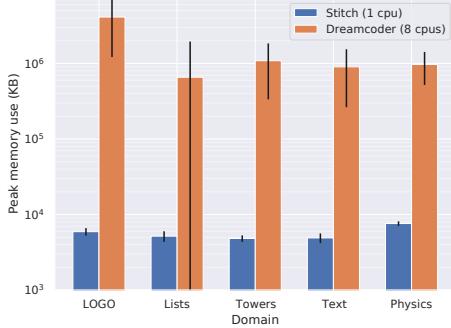


Fig. 8. Peak memory usage of STITCH and DreamCoder while running on the five domains considered in 6.1, averaged over all benchmarks. Lower is better; black lines indicate  $\pm$  one standard deviation. Note the logarithmic y-axis.

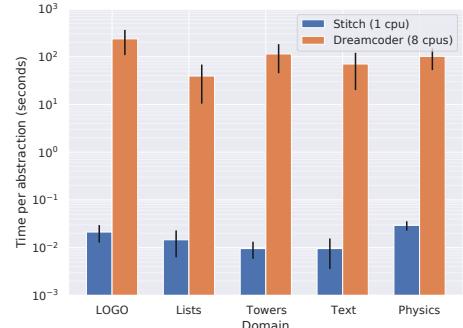


Fig. 9. Wall-clock time required to find (and rewrite under) one abstraction in each of the five domains from 6.1, averaged over all benchmarks. Lower is better; black lines indicate  $\pm$  one standard deviation. Note the logarithmic y-axis.

STITCH capable of handling most higher-order functions in Section 6.5, based on the formalism developed in Section 5. Nonetheless, we conclude that **STITCH learns libraries whose quality is comparable to and often better than those found by DreamCoder**.

**Resource efficiency.** In addition to the quality of the libraries found, we are interested in how the two methods compare in terms of time and space requirements. Since we a priori believe STITCH to be significantly faster, for this evaluation we allow DreamCoder to use 8 CPUs but limit STITCH to single threading<sup>5</sup>. The results are shown in Fig. 8 and Fig. 9; in summary, we find that STITCH takes *tens of milliseconds* to discover abstractions across all five domains—achieving a *3-4 order of magnitude* speed-up over DreamCoder—while also requiring more than *2 orders of magnitude* less memory. We thus conclude that **STITCH is dramatically more efficient than the state-of-the-art deductive baseline** when replicating the iterative library learning experiments of Ellis et al. 2021.

## 6.2 Large-scale Corpus Library Learning

**Experimental setup.** While the previous experiment allowed us to benchmark STITCH directly against a state-of-the-art deductive baseline, the iterated learning setting considered by Ellis et al. 2021 only evaluates library learning on relatively small corpora of short programs discovered by the synthesizer. Our second experiment is instead designed to evaluate STITCH in a more traditional learning setting, in which we aim to learn libraries of abstractions from a large corpus of existing programs all at once.

We source our larger-scale program datasets from Wong et al. 2022, which present a series of datasets designed as a benchmark for comparing human-level abstraction learning and graphics program writing against automated synthesis and library learning models. These datasets are divided into two distinct high-level domains (*technical drawings* and *block-tower planning tasks*), each consisting of four distinct subdomains containing 250 programs:

- **Technical drawing domains: nuts and bolts; vehicles; gadgets; furniture:** CAD-like graphics programs that render technical drawings of common objects, written in an initial DSL

<sup>5</sup>While this may seem unfair to STITCH, it is worth noting that it would be unlikely to benefit from multithreading when running on the order of milliseconds anyway; DreamCoder, on the other hand, would struggle greatly in these domains without the aid of parallelism.

consisting of looped transformations (scaling, translation, rotation) over simple geometric curves (lines and arcs).

- **Tower construction domains:** *bridges; cities; houses; castles*: Planning programs that construct complex architectures by placing blocks, written in an initial DSL that moves a virtual hand over a canvas and places horizontal and vertical bricks.

We choose these datasets not only for their size and scale (full dataset statistics in Table 1), but also for the complexity of their potential abstractions: Wong et al. 2022 explicitly design their corpora to contain complex hierarchical structures throughout the programs, making them an interesting setting for library learning.

Domain	#Programs	Average program length	Average program depth
nuts & bolts	250	$76.03 \pm 24.22$	$15.18 \pm 2.13$
gadgets	250	$142.85 \pm 87.32$	$20.88 \pm 2.37$
furniture	250	$171.74 \pm 48.41$	$31.83 \pm 5.33$
vehicles	250	$141.70 \pm 40.23$	$21.22 \pm 1.35$
bridges	250	$137.03 \pm 59.71$	$92.35 \pm 39.80$
cities	250	$161.70 \pm 55.56$	$109.80 \pm 37.66$
castles	250	$189.09 \pm 60.18$	$128.27 \pm 40.77$
houses	250	$168.13 \pm 55.75$	$114.85 \pm 37.60$

Table 1. Summary statistics about the domains from Wong et al. 2022. Program length is the number of terminal symbols in the program; program depth is the length of the longest path from root to leaf in the program tree. Both are reported as the mean over the entire dataset  $\pm$  one standard deviation.

When performing library learning in a synthesis setting by compressing a corpus of solutions, it’s desirable to find abstractions that would be useful for solving *new* tasks, as opposed to abstractions that overfit to the existing solutions. To evaluate how well the abstractions we learn apply to heldout programs in the domain, we split the corpora into train and test sets, running STITCH on the train set and evaluating its compression on the test set. Since Wong et al. 2022 do not present a train/test split of their datasets, we use stochastic cross validation to evaluate the generalization of the libraries found by STITCH. For each domain, we randomly sample 80% of the dataset to train on and reserve the last 20% as a held out test set; we repeat this procedure 50 times. We then ask STITCH to learn a library consisting of  $\leq 10$  abstractions with a maximum arity of 3, and average the results across the different random seeds.

To obtain a baseline to compare its performance against, we once again turned to DreamCoder [Ellis et al. 2021]. However, we found that DreamCoder was unable to discover even a single abstraction when run directly on any of the datasets from Wong et al. 2022, despite being given hours of runtime and 256GB of RAM. We also experimented with heavily sub-sampling the training dataset before passing it to DreamCoder, but failed to find a configuration under which DreamCoder finds any interesting abstractions at all due to the fact that it immediately blows up on programs as long as these. As a result, we resort to presenting STITCH’s performance metrics without any baseline to compare against; we stress that this is a direct result of the fact that STITCH is the first library learning tool that scales to such a challenging setting.

Domain	Compression Ratio		Runtime (s)	Peak mem. usage (MB)
	Training set	Test set		
nuts & bolts	12.00 ± 0.25	11.57 ± 0.49	0.24 ± 0.04	11.10 ± 0.17
gadgets	4.03 ± 0.15	3.91 ± 0.31	1.83 ± 0.35	30.11 ± 0.85
furniture	4.95 ± 0.07	4.85 ± 0.26	2.83 ± 0.49	30.88 ± 0.62
vehicles	4.28 ± 0.14	4.14 ± 0.25	1.52 ± 0.29	26.98 ± 0.88
bridges	4.36 ± 0.06	3.78 ± 0.14	17.58 ± 1.26	189.08 ± 10.10
cities	3.15 ± 0.05	3.06 ± 0.14	50.74 ± 3.96	413.63 ± 9.51
castles	2.57 ± 0.07	2.55 ± 0.08	77.26 ± 6.98	683.89 ± 32.96
houses	8.92 ± 0.21	8.85 ± 0.57	15.54 ± 1.37	241.17 ± 2.86

Table 2. Results for the large-scale library learning experiment in Sec. 6.2. The compression ratio refers to how many times smaller the corpus is after rewriting under the learned library compared to the original corpus; higher is thus better. All results are given as the mean ± one standard deviation over 50 runs with different random seeds for the dataset splitting.

*Results.* The results are summarized in Table 2. We find that STITCH scales up to even the most complex sub-domains, running in 77 seconds with a peak memory usage below 1GB on *castles*. On four out of eight of the domains, STITCH finishes in single-digit seconds and consumes only tens of megabytes. This stands in stark contrast to DreamCoder, which we were unable to run on the very simple nuts & bolts domain even with 256GB RAM and several hours worth of compute budget. These results thus support our claim that **STITCH scales to corpora of programs that would be intractable with prior library learning approaches**.

We hope that by providing our results on these datasets in full, future work in this field will benefit from having a directly comparable baseline.

### 6.3 Robustness to Early Search Termination

*Experimental setup.* This experiment is designed to evaluate how early into the search procedure STITCH finds what will eventually prove to be the optimal abstraction. This is highly relevant in settings where the set of training programs is too large to run the search to completion. The experiment showcases one of STITCH’s more subtle strengths: corpus-guided top-down abstraction search is an *anytime* algorithm, and thus does not need to be run to completion to give useful results.

We re-use the domains from Wong et al. 2022 and once again evaluate the quality of the library learned (measured in program compression), similarly to what was done in the previous experiment. However, since we are interested in how quickly STITCH finds *what it perceives to be* the optimal abstraction, we measure compressivity of the training dataset itself (rather than a held-out test set) and capture the compression ratio obtained by each *candidate* abstraction found *during* search (rather than just the compression ratio obtained when search has been run to completion). Thus, we are able to investigate how early on during the search procedure STITCH converges on a chosen library. We restrict STITCH to learning a single abstraction with a maximum arity of 3.

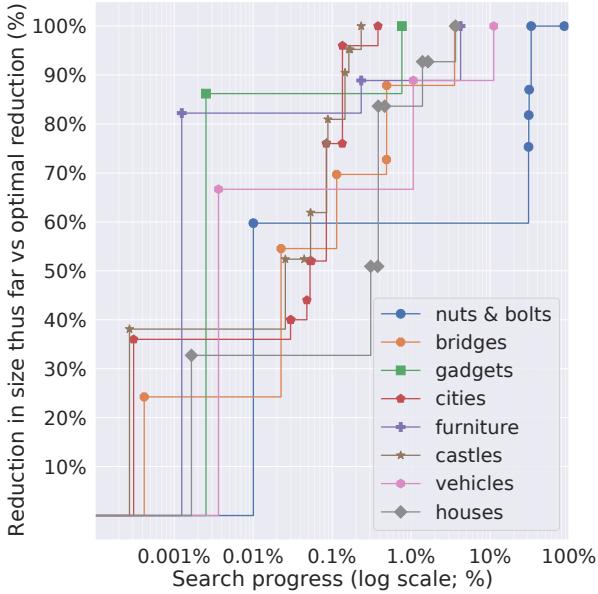


Fig. 10. Reduction in size obtained on the training set when rewritten under the best abstraction found thus far vs. the number of nodes expanded during search. The y-axis is normalized with respect to the optimal abstraction; the x-axis is normalized with respect to the total number of nodes explored by STITCH. Lines thus end earlier the quicker STITCH finds the optimal abstraction; however in all runs STITCH *continues to search* until reaching 100% on the x-axis, exploring the rest of the abstraction space without finding a new best abstraction. Note the logarithmic x-axis.

**Results.** The results are shown in Fig. 10. These results validate our hypothesis that **STITCH is empirically robust to terminating the search procedure early**: in every sub-domain except for *nuts & bolts* STITCH converges to the optimal abstraction very early on, having only completed a tiny fraction of the total search.<sup>6</sup> We believe that this has great importance for STITCH’s applicability in data-rich settings since it suggests that a nearly-optimal abstraction can be found even if the search must be terminated early (e.g. after a fixed amount of time has passed), making early stopping an empirically useful way of speeding up the library learning process.

#### 6.4 Ablation Study

**Experimental setup.** STITCH implements several different optimizations, which we have argued hasten the search for abstractions. To verify that this claim holds in practice, we now carry out a brief ablation study.

Since the space of every possible combination of optimizations is too large to present succinctly, we focus our attention on four ablations:

- **no-arg-capture** (from Section 4.3), which disables the pruning of abstractions which are only ever used with the exact same set of arguments (and these arguments could therefore just be in-lined for greater compression).
- **no-upper-bound**, which disables the upper bound based pruning.

<sup>6</sup>Given that significant attention has already been given to wall-clock run-times of STITCH on similar workloads in 6.2, we here use the number of nodes explored instead of wall-clock time to ensure deterministic and easily reproducible results.

System \ Domain	bridges	castles	cities	gadgets	furniture	houses	nuts & bolts	vehicles
System								
no-arg-capture	RAM	RAM	RAM	16714.28	RAM	RAM	30.21	26.67
no-upper-bound	12.15	23.71	27.90	207.76	119.00	36.18	179.12	179.99
no-redundant-args	1.95	1.42	1.37	1.01	1.00	1.01	1.09	1.00
no-opt	RAM	RAM	RAM	⌚	RAM	RAM	⌚	RAM

Table 3. Results from the ablation study. Each cell contains the ratio between the number of nodes explored during search by that particular system and the number of nodes explored by the baseline on the same domain. Lower is better; 1.00 means performance is identical to the baseline. Cells labeled ~~RAM~~ crashed due to reaching the 50GB virtual address space limit, while those labeled ⌚ reached the 90 minute time restriction.

- **no-redundant-args** (from Section 4.3), which disables the pruning of multi-argument abstractions that have a redundant argument that could be removed because it is always the same as another argument.
- **no-opt**, which disables all of STITCH’s optimizations.

To isolate the impact of disabling optimizations, we run a single iteration of abstraction learning on each of the 8 domains from 6.2 and collect the number of nodes explored during the search. The first iteration is generally the most challenging as the corpus is large and has not yet been compressed at all. Focusing on the number of nodes explored (rather than for example runtime) allows for deterministic results. We also fix the maximum arity of abstractions to 3 in all runs, aiming to strike a good balance between compute requirements and how much the optimization will be exposed. We limit each run to 50GB of virtual address space, as well as 90 minutes of compute.

*Results.* The results are shown in Table 3. We first note as a sanity check that each ablation does indeed lead to reduced performance in general (i.e. explores a search space larger than the baseline does). Furthermore, the results suggest that upper bound based pruning is the most important in the *nuts & bolts* and *vehicles* domains, while pruning out argument capture abstractions is the most important in the other six domains; it is noteworthy that this latter ablation by itself causes STITCH to reach the memory limit on more than half of the domains. On the other hand, disabling the pruning of redundant arguments has a relatively modest impact on the size of the search space, but still leads to an almost 2x improvement in the bridge domain.

Perhaps the most important takeaway from this ablation study is that when *all* optimizations are disabled, STITCH fails to find an abstraction within the resource budget on any of the domains. This verifies our hypothesis that corpus-guided pruning of the search space is the key factor involved in making top-down synthesis of abstractions tractable.

## 6.5 Learning Libraries of Higher-Order Functions

*Experimental setup.* Our experiments up till now have focused on performant and scalable library learning. This comes at the expense of some expressivity: deductive rewrite systems can, in principle, express broader spaces of refactorings. For example, a rewrite based on inverting  $\beta$ -reduction allows inventing auxiliary  $\lambda$ -abstractions, which helps with learning higher-order functions: in Ellis et al. 2021, DreamCoder is shown to recover higher-order functions such as `map`, `fold`, `unfold`, `filter`,

and `zip_with`, starting from the Y-combinator. This works by constructing a version space which encodes every refactoring that is equivalent up to  $\beta$ -inversion rewrites. But DreamCoder’s coverage comes at a steep cost as inverting  $\beta$ -reduction is expensive.

In this experiment, we seek to give evidence that it is possible to make STITCH recover all of these higher-order functions by layering it on top of the version space obtained after a single step of DreamCoder’s  $\beta$ -inversion, following the formalism outlined in Section 5. We then compare this modified version of STITCH with DreamCoder on its ability to learn these higher-order functions from programs generated by intermediate DreamCoder iterations, and measure the runtime of each approach.

*Probabilistic Re-ranking.* While the approach outlined in Section 5 should suffice to layer STITCH on top of deductive rewrite systems based on version spaces, some extra care needs to be taken to combine it with DreamCoder. This is because DreamCoder implements a probabilistic Bayesian objective for judging candidate abstractions (exploiting the connection between compression and probability [Shannon 1948]), seeking the library  $L$  which maximizes  $P(L) \prod_{p \in \mathcal{P}} P(p|L)$  for a given or learned prior  $P(L)$  and program likelihood  $P(p|L)$ . STITCH, on the other hand, effectively judges compression quality via a cost function capturing the (weighted) size of the programs as detailed in Section 4.

To implement this probabilistic heuristic in STITCH, we simply run STITCH on the version spaces as-is but then re-score each complete abstraction popped off of the priority queue under the Bayesian objective, using DreamCoder’s models of the prior and likelihood. To make this integration easier, we re-implement STITCH in Python, giving a prototype version called `pySTITCH` which is only used for this experiment. Our implementation employs strict dominance pruning, and prunes using the bound on the approximate utility given in Section 5.

System	Fold	Unfold	Map	Filter	ZipWith	Time (s)
<code>pySTITCH</code>	Base	✓	✗	✓	✗	503
	+Bayes	✓	✗	✓	✗	817
	+VS	✓	✓	✓	✗	3231
	+Bayes+VS	✓	✓	✓	✓	3042
DreamCoder	Step 1	✗	✗	✓	✗	67
	Step 2	✗	✗	✓	✓	116
	Step 3	✗	✗	✓	✓	2254
	Step 4	✓	✓	✓	✓	228048

Table 4. Comparing library learners on functional programming exercises. DreamCoder,  $n$ -steps: deductive baseline rewriting  $n$  steps of  $\beta$ -reduction. `pySTITCH`: Python reimplementation of STITCH, which enables better interoperability with DreamCoder’s version space algebra (+VS) and probabilistic models (+Bayes). `pySTITCH+VS+Bayes` learns all of the same higher-order functions, using < 2% of the compute.

*Results.* The results are shown in Table 4. We note first the large discrepancy in runtimes; running DreamCoder with 4 steps of rewriting takes roughly 2.5 days of compute on this domain, while even the slowest version of `pySTITCH` still finishes in less than an hour. In terms of the number of

higher-order abstractions found, DreamCoder only finds all five when run in its most expensive configuration; reducing the computation cost quickly decreases its expressivity. For `PYSTITCH`, the Bayesian re-ranking alone (`PYSTITCH+Bayes`) does not yield any improvements, while running it on the version spaces (`PYSTITCH+VS`) yields 4 out of 5 functions. However, it is only when these two adaptations are used in conjunction (`PYSTITCH+Bayes+VS`) that the `STITCH`-based method is able to find all of the higher-order functions.

In summary, we find that running CTS after a single step of version space rewriting and then probabilistically re-ranking the results suffices to recover the core higher-order functions that DreamCoder learns, while using <2% of its compute. We thus conclude that **running `STITCH` on top of a deductive rewrite system reduces the expressivity gap, while retaining superior performance.**

## 7 RELATED WORK

`STITCH` is related to two core ideas from prior work: **deductive refactoring and library learning systems**, which introduce the idea of learning abstractions that capture common structure across a set of programs, but have largely been driven by deductive algorithms; and **guided top-down program synthesis systems**, which use cost functions to guide top-down enumerative search over a space of programs, but have largely been used to synthesize whole programs for individual tasks in prior work.

### 7.1 Deductive Refactoring and Library Learning

Recent work shares `STITCH`'s goal of learning *libraries* of program abstractions which capture reusable structure across a corpus of programs [Allamanis and Sutton 2014; Cropper 2019; Dechter et al. 2013; Ellis et al. 2021, 2018; Iyer et al. 2019; Jones et al. 2021; Shin et al. 2019; Wong et al. 2021]. Several of these prior approaches also introduce a utility metric based on *program compression* in order to determine the most useful candidate abstractions to retain [Dechter et al. 2013; Ellis et al. 2021; Iyer et al. 2019; Lázaro-Gredilla et al. 2019; Wong et al. 2021].

Much of this prior work follows a bottom-up approach to abstraction learning, combining a bottom-up traversal across individual training programs with a second stage to extract shared abstractions from across the training corpus. This approach includes systems that work through direct *memoization* of subtrees across a corpus of programs [Dechter et al. 2013; Lázaro-Gredilla et al. 2019; Lin et al. 2014]; *antiunification* (caching tree templates that can be unified with training program syntax trees) [Ellis et al. 2018; Henderson 2013; Hwang et al. 2011; Iyer et al. 2019]; or by more sophisticated *refactoring* using one or more rewrite rules to expose additional shared structure across training programs [Chlipala et al. 2017; Ellis et al. 2021, 2018; Liang et al. 2010].

Many of the bottom up algorithms draw more generally on *deductive synthesis* approaches that apply local rewrite rules in a bottom-up fashion to program trees in order to refactor them – historically, to synthesize programs from a declarative specification of desired function [Burstall and Darlington 1977; Manna and Waldinger 1980]. Deductive approaches to library learning, however, confront fundamental memory and search-time scaling challenges as the corpus size and depth of the training programs increases; prior deductive approaches such as [Ellis et al. 2021, 2018] use version spaces [Lau et al. 2003; Mitchell 1977] to mitigate the memory usage during bottom-up abstraction proposal. Still, deductive approaches are challenging to bound and prune (unlike the top-down approach we take in `STITCH`), as they generally traverse individual program trees locally and must store possible abstraction candidates in memory before the extraction step.

Some prior work [Allamanis and Sutton 2014; Shin et al. 2019] takes an MCMC-based approach with better scaling behavior than deductive rewrites, but differs from the goals of `STITCH` as they don't address binders and focus on common syntactic fragments instead of well-formed functions.

## 7.2 Guided Top-Down Program Synthesis

STITCH uses a corpus-guided top-down approach to learning library abstractions that is closely related to recent *guided enumerative synthesis techniques*. This includes methods that leverage type-based constraints on holes [Feser et al. 2015; Polikarpova et al. 2016], over- and under-approximations of the behaviors of holes [Chen et al. 2020; Lee et al. 2016], and probabilistic techniques to heuristically guide the search [Balog et al. 2016; Ellis et al. 2020, 2021; Nye et al. 2021; Shah et al. 2020]. These approaches have largely been applied in to synthesize entire programs based on input/output examples or another form of specification, in contrast to the abstraction-learning goal in our work [Allamanis et al. 2018; Balog et al. 2016; Chen et al. 2018; Ellis et al. 2018; Ganin et al. 2018; Koukoutos et al. 2017]. Like STITCH however, these approaches sometimes use cost functions (such as the likelihood of a partially enumerated program under a hand-crafted or learned probabilistic generative model over programs) in order to direct search towards more desirable program trees. However, STITCH’s cost function leverages a more direct relationship between partially-enumerated candidate functions and the existing training corpus, unlike the cost functions typically applied in inductive synthesis, which must be estimated from input/output examples.

## 7.3 Lambda-aware unification

The LAMBDAUNIFY procedure presented in Section 3 relates to prior work on unification modulo binders [Dowek et al. 1995, 1996; Huet 1975; Miller 1991, 1992].

*Notion of beta-equivalence.* This prior work is concerned with more general notions of equivalence modulo beta-reduction, while LAMBDAUNIFY is based on a restricted but fast syntax-driven equivalence. For example, in Dowek et al. 1996 one might try to unify  $(\alpha \text{ foo})$  with  $(\text{foo})$  and get the two solutions:  $[\alpha \rightarrow (\lambda. \$0)]$  and  $[\alpha \rightarrow (\lambda. \text{foo})]$ . In contrast, LAMBDAUNIFY will never introduce a  $\lambda$ -abstraction to create a higher order argument and thus  $(\alpha \text{ foo})$  and  $(\text{foo})$  would not unify at all, since one is an application while the other is a primitive. Instead, higher order abstraction arguments are handled by combining the algorithm with a deductive approach when desirable as in Section 5. We also note that in LAMBDAUNIFY( $A, e$ ) it is assumed that  $e$  is an expression and thus does not contain abstraction variables, which further simplifies the approach compared to this prior work.

*Handling of binders.* Dowek et al. 1995 and Dowek et al. 1996 both employ de Bruijn indexed variables and therefore must similarly account for the shifting of variables in arguments when inverting beta reduction. While LAMBDAUNIFY handles this through DOWNSHIFTALL, Dowek et al. 1995 and Dowek et al. 1996 both converting their terms to the  $\lambda\sigma$ -calculus of explicit substitutions [Abadi et al. 1989] which allows them to insert upshifting operators at each abstraction variable location. This alternate handling of shifting is useful given the more general notion of equivalence they are considering, but is excessive for our simpler syntax-guided task.

*Handling of holes.* In LAMBDAUNIFY( $A, e$ ) we allow  $A$  to contain holes  $??$  which are allowed to violate index shifting rules during unification as they are considered unfinished subtrees of the abstraction. This is handled through the use of  $\&i$  indices. While the special-case handling of holes in expressions is not directly part of this prior work, there are similarities between it and the Skolemization [Skolem 1920] done in Miller 1992. Skolemization allows for lifting an existential quantifier (i.e. an abstraction variable or hole) above a universal quantifier (i.e. a lambda) by turning the abstraction variable or hole into a function of its local context – in essence piping the local context into the hole. In the context of the lambda calculus this is essentially a form of lambda lifting [Johnsson 1985], which is the process of lifting a local function that contains free variables by binding the free variables as additional arguments to the function and passing them in at each call site. In Miller 1992 this is used for abstraction variables (as there are no holes) and aids in

the more general notion of beta-equivalence they are dealing with, while for our purposes the  $\&$  variables suffice and don't require extra manipulations of lambdas to thread in the local context.

#### 7.4 Upper Bounds in Network Motifs

There is also an interesting connection between this work and prior work on finding *network motifs* [Milo et al. 2002], which are frequently-occurring subgraphs within a corpus of graphs. One approach to mining high-frequency motifs is based on growing a motif one edge at a time, much like we grow abstractions one node at a time during our search [Kuramochi and Karypis 2001, 2004; Schreiber and Schwöbbermeyer 2005]. This approach makes use of the insight that the frequency of a motif will only decrease as it is grown, giving an upper bound on the frequency of any motif derived from it. While we employ a more complex compressive utility function than their frequency-based utility, our upper bound is based on a similar insight that larger abstractions will match at a subset of the locations.

#### 7.5 Comparison to BABBLE

BABBLE [Cao et al. 2023] is concurrent work in library learning that likewise adopts the compression objective from DreamCoder. While BABBLE focuses primarily on expressive library learning through an algorithm that can reason over semantic equivalences represented through e-graphs, STITCH focuses primarily on efficient library learning through a parallel anytime branch-and-bound algorithm.

STITCH synthesizes the maximally-compressive abstraction for a corpus of programs through a branch-and-bound top-down search, computing an upper bound on the compression of any partially-constructed abstraction to guide and prune the search. In contrast, BABBLE operates on an e-graph representing many equivalent corpora and uses anti-unification to propose candidate abstractions and a custom e-graph beam-search extraction algorithm to select a maximally-compressive set of abstractions from the proposal.

At a high level, STITCH aims to be *efficient* while BABBLE aims to be *expressive*. The key differences between the algorithms are:

- *Efficiency*: On DreamCoder's benchmarks, BABBLE is 10-100x faster than DreamCoder while STITCH is 1,000-10,000x faster than DreamCoder. Unlike BABBLE, STITCH is an anytime algorithm that can be stopped early for a best-so-far result.
- *Expressivity*: BABBLE can learn libraries modulo equational theories, allowing it to find common semantic abstractions despite syntactic differences in the programs. STITCH only explores non-syntactic equivalences preliminarily in the PySTITCH prototype with the DreamCoder beta-inversion rewrite rule (Section 5 and Section 6.5).
- *Jointly learning a library*: BABBLE can learn multiple abstractions at once that jointly provide compression, while DreamCoder and STITCH repeatedly learn a single abstraction at a time. STITCH provides an optimality guarantee on each abstraction learned while BABBLE approximates the joint objective through a beam search.

These two approaches have advantages and disadvantages, and we believe that there is strong potential in combining the two.

### 8 CONCLUSION AND FUTURE WORK

We have presented *corpus-guided top-down synthesis* (CTS)—an efficient new algorithm for synthesizing libraries of functional abstractions capturing common functionality within a corpus of programs. CTS directly synthesizes the abstractions, rather than exposing them through a series of

rewrites as is done by deductive systems. Key to its performance is the usage of a guiding utility function, which allows CTS to effectively search (and prune out large portions of) the space of possible abstractions.

We implement this algorithm in STITCH, an open-source library learning tool which exposes both Rust and Python bindings<sup>7</sup>. We evaluate STITCH across five experimental settings, demonstrating that it learns comparably compressive libraries with 2 orders of magnitude less memory and 3-4 orders of magnitude less time compared to the state-of-the-art deductive algorithm of Ellis et al. 2021. We also find that STITCH scales to learning libraries of abstractions from much larger datasets of deeper program trees than is possible with prior work, and that the *anytime* property of corpus-guided top-down search — abstractions discovered via top-down search are already compressive early in search and improve as it continues — offers the opportunity for high-quality library learning even in complex domains through early search termination.

There remain many open problems and exciting directions in abstraction learning. One major direction is extending the CTS approach to handle reasoning over more general deductive rewrite systems, like those encoded in egg [Willsey et al. 2021] as BABBLE does [Cao et al. 2023] (also see Section 7.5). The PySTITCH prototype presented in Section 5 only works with a single rewrite rule (beta-inversion) and uses an approximate utility without an optimality guarantee. This is because it is challenging to adapt the upper bound used by STITCH to version spaces or e-graphs due to the trade-off between the compression gained from equational rewrites and compression gained from abstraction learning.

The speedup afforded by STITCH may allow for new applications of abstraction learning as well, as it can now be used as an inexpensive subroutine that can be called much more frequently than prior algorithms. To further extend its range of applications, CTS could also be adapted to finding abstractions in data structures beyond program trees, such as dataflow DAGs and more general graph structures like molecules, if the bounds and matching can be adapted. Finally, while we work with fairly large lambda calculus programs in this work, there is a clear gap in scale between these programs and those in an actual codebase, so exploring applications of CTS to real-world code is an exciting direction.

## ACKNOWLEDGMENTS

We thank A. Lew and J. Andreas for helpful discussions, and J. Feser and I. Kuraj for feedback on the manuscript. M.B. and G.G are supported by the National Science Foundation (NSF) Graduate Research Fellowship under Grant No. 2141064. M.B. is also supported by the Defense Advanced Research Projects Agency (DARPA) under the SDCPS Contract FA8750-20-C-0542. T.X.O. is supported by Herbert E. Grier (1933) and Dorothy J. Grier Fund Fellowship. C.W. and J.B.T. are supported by AFOSR under grant number FA9550-19-1-0269, the MIT Quest for Intelligence, the MIT-IBM Watson AI Lab, ONR Science of AI, and DARPA Machine Common Sense. G.G. is also supported by the MIT Presidential Fellowship. A.S. is supported by the NSF under Grant No. 1918839. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of sponsors.

---

<sup>7</sup><https://github.com/mlb2251/stitch>

## REFERENCES

- Martin Abadi, Luca Cardelli, P-L Curien, and J-J Lévy. 1989. Explicit substitutions. In *Proceedings of the 17th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*. 31–46.
- Miltiadis Allamanis, Earl T Barr, Premkumar Devanbu, and Charles Sutton. 2018. A survey of machine learning for big code and naturalness. *ACM Computing Surveys (CSUR)* 51, 4 (2018), 1–37.
- Miltiadis Allamanis and Charles Sutton. 2014. Mining idioms from source code. In *Proceedings of the 22nd ACM SIGSOFT international symposium on foundations of software engineering*. 472–483.
- Matej Balog, Alexander L Gaunt, Marc Brockschmidt, Sebastian Nowozin, and Daniel Tarlow. 2016. Deepcoder: Learning to write programs. *arXiv preprint arXiv:1611.01989* (2016).
- Rod M Burstall and John Darlington. 1977. A transformation system for developing recursive programs. *Journal of the ACM (JACM)* 24, 1 (1977), 44–67.
- David Cao, Rose Kunkel, Chandrakana Nandi, Max Willsey, Zachary Tatlock, and Nadia Polikarpova. 2023. babble: Learning Better Abstractions with E-Graphs and Anti-Unification. *Proc. ACM Program. Lang. POPL* (2023). <https://doi.org/10.1145/3571207>
- Qiaochu Chen, Xinyu Wang, Xi Ye, Greg Durrett, and Isil Dillig. 2020. Multi-modal synthesis of regular expressions. In *Proceedings of the 41st ACM SIGPLAN International Conference on Programming Language Design and Implementation, PLDI 2020, London, UK, June 15–20, 2020*, Alastair F. Donaldson and Emina Torlak (Eds.). ACM, 487–502. <https://doi.org/10.1145/3385412.3385988>
- Xinyun Chen, Chang Liu, and Dawn Song. 2018. Execution-guided neural program synthesis. In *International Conference on Learning Representations*.
- Adam Chlipala, Benjamin Delaware, Samuel Duchovni, Jason Gross, Clément Pit-Claudel, Sorawit Suriyakarn, Peng Wang, and Katherine Ye. 2017. The end of history? Using a proof assistant to replace language design with library design. In *2nd Summit on Advances in Programming Languages (SNAPL 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.
- Geoffrey Chu and Peter J. Stuckey. 2015. Dominance breaking constraints. *Constraints An Int. J.* 20, 2 (2015), 155–182. <https://doi.org/10.1007/s10601-014-9173-7>
- Andrew Cropper. 2019. Playgol: Learning programs through play. *arXiv preprint arXiv:1904.08993* (2019).
- Nicolaas Govert de Bruijn. 1972. Lambda calculus notation with nameless dummies, a tool for automatic formula manipulation, with application to the Church-Rosser theorem. In *Indagationes Mathematicae (Proceedings)*, Vol. 75. Elsevier, 381–392.
- Eyal Dechter, Jonathan Malmaud, Ryan Prescott Adams, and Joshua B Tenenbaum. 2013. Bootstrap learning via modular concept discovery. In *Proceedings of the International Joint Conference on Artificial Intelligence*. AAAI Press/International Joint Conferences on Artificial Intelligence.
- Gilles Dowek, Thérèse Hardin, and Claude Kirchner. 1995. Higher-Order Unification via Explicit Substitutions. In *Proceedings of the Tenth Annual Symposium on Logic in Computer Science*, D. Kozen (Ed.). IEEE Computer Society Press, San Diego, California, 366–374.
- Gilles Dowek, Thérèse Hardin, Claude Kirchner, and Frank Pfenning. 1996. Unification via Explicit Substitutions: The Case of Higher-Order Patterns. In *Proceedings of the Joint International Conference and Symposium on Logic Programming*, M. Maher (Ed.). MIT Press, Bonn, Germany, 259–273.
- Kevin Ellis, Catherine Wong, Maxwell Nye, Mathias Sablé-Meyer, Luc Cary, Lucas Morales, Luke Hewitt, Armando Solar-Lezama, and Joshua B Tenenbaum. 2020. Dreamcoder: Growing generalizable, interpretable knowledge with wake-sleep bayesian program learning. *arXiv preprint arXiv:2006.08381* (2020).
- Kevin Ellis, Catherine Wong, Maxwell Nye, Mathias Sablé-Meyer, Lucas Morales, Luke Hewitt, Luc Cary, Armando Solar-Lezama, and Joshua B Tenenbaum. 2021. Dreamcoder: Bootstrapping inductive program synthesis with wake-sleep library learning. In *Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation*. 835–850.
- Kevin M Ellis, Lucas E Morales, Mathias Sablé-Meyer, Armando Solar Lezama, and Joshua B Tenenbaum. 2018. Library learning for neurally-guided bayesian program induction. (2018).
- Matthias Felleisen and Robert Hieb. 1992. The Revised Report on the Syntactic Theories of Sequential Control and State. *Theor. Comput. Sci.* 103, 2 (sep 1992), 235–271. [https://doi.org/10.1016/0304-3975\(92\)90014-7](https://doi.org/10.1016/0304-3975(92)90014-7)
- John K Feser, Swarat Chaudhuri, and Isil Dillig. 2015. Synthesizing data structure transformations from input-output examples. *ACM SIGPLAN Notices* 50, 6 (2015), 229–239.
- Yaroslav Ganin, Tejas Kulkarni, Igor Babuschkin, SM Ali Eslami, and Oriol Vinyals. 2018. Synthesizing programs for images using reinforced adversarial learning. In *International Conference on Machine Learning*. PMLR, 1666–1675.
- Sumit Gulwani, José Hernández-Orallo, Emanuel Kitzelmann, Stephen H Muggleton, Ute Schmid, and Benjamin Zorn. 2015. Inductive programming meets the real world. *Commun. ACM* 58, 11 (2015), 90–99.
- Peter E. Hart, Nils J. Nilsson, and Bertram Raphael. 1968. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Trans. Syst. Sci. Cybern.* 4, 2 (1968), 100–107. <https://doi.org/10.1109/TSSC.1968.300136>

- Robert John Henderson. 2013. Cumulative learning in the lambda calculus. (2013).
- Gérard Huet. 1975. A Unification Algorithm for Typed  $\lambda$ -Calculus. *Theoretical Computer Science* 1 (1975), 27–57.
- Irvin Hwang, Andreas Stuhlmüller, and Noah D Goodman. 2011. Inducing probabilistic programs by Bayesian program merging. *arXiv preprint arXiv:1110.5667* (2011).
- Toshihide Ibaraki. 1977. The Power of Dominance Relations in Branch-and-Bound Algorithms. *J. ACM* 24, 2 (1977), 264–279. <https://doi.org/10.1145/322003.322010>
- Srinivasan Iyer, Alvin Cheung, and Luke Zettlemoyer. 2019. Learning programmatic idioms for scalable semantic parsing. *arXiv preprint arXiv:1904.09086* (2019).
- Thomas Johnsson. 1985. Lambda Lifting: Transforming Programs to Recursive Equations. In *Functional Programming Languages and Computer Architecture, FPCA 1985, Nancy, France, September 16–19, 1985, Proceedings (Lecture Notes in Computer Science, Vol. 201)*, Jean-Pierre Jouannaud (Ed.). Springer, 190–203. [https://doi.org/10.1007/3-540-15975-4\\_37](https://doi.org/10.1007/3-540-15975-4_37)
- R Kenny Jones, David Charatan, Paul Guerrero, Niloy J Mitra, and Daniel Ritchie. 2021. ShapeMOD: macro operation discovery for 3D shape programs. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–16.
- Manos Koukoutos, Mukund Raghothaman, Etienne Kneuss, and Viktor Kuncak. 2017. On repair with probabilistic attribute grammars. *arXiv preprint arXiv:1707.04148* (2017).
- Michihiro Kuramochi and George Karypis. 2001. Frequent Subgraph Discovery. In *Proceedings of the 2001 IEEE International Conference on Data Mining, 29 November - 2 December 2001, San Jose, California, USA*, Nick Cercone, Tsau Young Lin, and Xindong Wu (Eds.). IEEE Computer Society, 313–320. <https://doi.org/10.1109/ICDM.2001.989534>
- Michihiro Kuramochi and George Karypis. 2004. Finding Frequent Patterns in a Large Sparse Graph. In *Proceedings of the Fourth SIAM International Conference on Data Mining, Lake Buena Vista, Florida, USA, April 22–24, 2004*, Michael W. Berry, Umeshwar Dayal, Chandrika Kamath, and David B. Skillicorn (Eds.). SIAM, 345–356. <https://doi.org/10.1137/1.9781611972740.32>
- A. H. Land and A. G. Doig. 1960. An Automatic Method of Solving Discrete Programming Problems. *Econometrica* 28, 3 (1960), 497–520. <http://www.jstor.org/stable/1910129>
- Tessa Lau, Steven A Wolfman, Pedro Domingos, and Daniel S Weld. 2003. Programming by demonstration using version space algebra. *Machine Learning* 53, 1 (2003), 111–156.
- Miguel Lázaro-Gredilla, Dianhuan Lin, J Swaroop Guntupalli, and Dileep George. 2019. Beyond imitation: Zero-shot task transfer on robots by learning concepts as cognitive programs. *Science Robotics* 4, 26 (2019), eaav3150.
- Mina Lee, Sunbeam So, and Hakjoo Oh. 2016. Synthesizing regular expressions from examples for introductory automata assignments. In *Proceedings of the 2016 ACM SIGPLAN International Conference on Generative Programming: Concepts and Experiences, GPCE 2016, Amsterdam, The Netherlands, October 31 - November 1, 2016*, Bernd Fischer and Ina Schaefer (Eds.). ACM, 70–80. <https://doi.org/10.1145/2993236.2993244>
- Percy Liang, Michael I Jordan, and Dan Klein. 2010. Learning programs: A hierarchical Bayesian approach. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 639–646.
- Dianhuan Lin, Eyal Dechter, Kevin Ellis, Joshua B Tenenbaum, and Stephen H Muggleton. 2014. Bias reformulation for one-shot function induction. (2014).
- Zohar Manna and Richard Waldinger. 1980. A deductive approach to program synthesis. *ACM Transactions on Programming Languages and Systems (TOPLAS)* 2, 1 (1980), 90–121.
- Dale Miller. 1991. A Logic Programming Language with Lambda-Abstraction, Function Variables, and Simple Unification. *Journal of Logic and Computation* 1, 4 (1991), 497–536.
- Dale Miller. 1992. Unification under a Mixed Prefix. *Journal of Symbolic Computation* 14 (1992), 321–358.
- R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. 2002. Network Motifs: Simple Building Blocks of Complex Networks. *Science* 298, 5594 (2002), 824–827. <https://doi.org/10.1126/science.298.5594.824> arXiv:<https://www.science.org/doi/pdf/10.1126/science.298.5594.824>
- Tom M Mitchell. 1977. Version spaces: A candidate elimination approach to rule learning. In *Proceedings of the 5th international joint conference on Artificial intelligence-Volume 1*, 305–310.
- David R. Morrison, Sheldon H. Jacobson, Jason J. Sauppe, and Edward C. Sewell. 2016. Branch-and-bound algorithms: A survey of recent advances in searching, branching, and pruning. *Discret. Optim.* 19 (2016), 79–102. <https://doi.org/10.1016/j.dsopt.2016.01.005>
- Maxwell Nye, Yewen Pu, Matthew Bowers, Jacob Andreas, Joshua B Tenenbaum, and Armando Solar-Lezama. 2021. Representing Partial Programs with Blended Abstract Semantics. In *International Conference on Learning Representations*.
- Nadia Polikarpova, Ivan Kuraj, and Armando Solar-Lezama. 2016. Program synthesis from polymorphic refinement types. *ACM SIGPLAN Notices* 51, 6 (2016), 522–538.
- Oleksandr Polozov and Sumit Gulwani. 2015. Flashmeta: A framework for inductive program synthesis. In *Proceedings of the 2015 ACM SIGPLAN International Conference on Object-Oriented Programming, Systems, Languages, and Applications*. 107–126.

- Falk Schreiber and Henning Schwöbbermeyer. 2005. Frequency Concepts and Pattern Detection for the Analysis of Motifs in Networks. *Trans. Comp. Sys. Biology* 3 (2005), 89–104. [https://doi.org/10.1007/11599128\\_7](https://doi.org/10.1007/11599128_7)
- Ameesh Shah, Eric Zhan, Jennifer Sun, Abhinav Verma, Yisong Yue, and Swarat Chaudhuri. 2020. Learning Differentiable Programs with Admissible Neural Heuristics. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 4940–4952. <https://proceedings.neurips.cc/paper/2020/file/342285bb2a8cadef22f667eeb6a63732-Paper.pdf>
- Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal* 27, 3 (1948), 379–423.
- Eui Chul Shin, Miltiadis Allamanis, Marc Brockschmidt, and Alex Polozov. 2019. Program synthesis and semantic parsing with learned code idioms. *Advances in Neural Information Processing Systems* 32 (2019).
- Thoralf Skolem. 1920. Logisch-kombinatorische Untersuchungen über die Erfüllbarkeit oder Beweisbarkeit mathematischer Sätze nebst einem Theorem über dichte Mengen. (1920).
- Max Willsey, Chandrakana Nandi, Yisu Remy Wang, Oliver Flatt, Zachary Tatlock, and Pavel Panchekha. 2021. Egg: Fast and Extensible Equality Saturation. *Proc. ACM Program. Lang.* 5, POPL, Article 23 (jan 2021), 29 pages. <https://doi.org/10.1145/3434304>
- Catherine Wong, Kevin M Ellis, Joshua Tenenbaum, and Jacob Andreas. 2021. Leveraging language to learn program abstractions and search heuristics. In *International Conference on Machine Learning*. PMLR, 11193–11204.
- Catherine Wong, William McCarthy, Gabriel Grand, Jacob Andreas, Joshua B Tenenbaum, Robert Hawkins, and Judy Fan. 2022. Identifying concept libraries from language about object structure. In *CogSci*. To appear.

## A CORPUS-GUIDED TOP-DOWN SEARCH: FULL ALGORITHM

In this appendix, we give an expansive description of the full corpus-guided top-down search (CTS) algorithm to ease re-implementation.

The full algorithm is given in Algorithm 1. It takes as input a corpus of programs  $\mathcal{P}$ , a utility function  $U_{\mathcal{P}, \mathcal{R}}(A)$ , and utility upper bound function  $\bar{U}_{\mathcal{P}, \mathcal{R}}(A_{??})$ . The algorithm searches for and returns the complete abstraction  $A_{best}$  that maximizes the utility function. The algorithm maintains a priority queue  $Q$  of partial abstractions ordered by their utility upper bound (or alternatively a custom priority function), a best complete abstraction so-far  $A_{best}$ , and a corresponding best utility so far  $U_{best}$  for that abstraction. The priority queue is initialized to hold the single-hole partial abstraction  $??$ , from which any abstraction can be derived.  $A_{best}$  and  $U_{best}$  are initialized as the best arity-zero abstraction and corresponding utility (line 2), since arity-zero abstractions can be quickly and completely enumerated (each unique, closed subexpression in the corpus is an arity-zero abstraction).

We then proceed to the core loop of the algorithm from lines 4-29. At each step of this loop we pop the highest-priority partial abstraction  $A_{??}$  off of the priority queue and process it. We discard  $A_{??}$  if its utility upper bound doesn't exceed our best utility found so far (lines 6-8). We then choose a hole  $h$  in  $A_{??}$  to expand with the procedure `CHOOSE-HOLE`. `CHOOSE-HOLE` can be a custom function; we find from preliminary experiments on a subset of the datasets that choosing the most recently introduced hole is effective in practice.

The algorithm then uses the procedure `EXPANSIONS` to iterate over all possible single step expansions of the hole  $h$  in  $A_{??}$ , such as replacing the hole with  $+$  or with  $(app ?? ??)$ . For each expanded abstraction  $A'_{??}$ , it is easy to compute its set of match locations  $MATCHES(\mathcal{P}, A'_{??})$  since we know this will be a subset of the match locations of  $A_{??}$  (and these will be *disjoint* subsets, except when expanding into an abstraction variable  $\alpha$ ). We can easily inspect the relevant subtree at each match location of the original abstraction to see which expansions are valid and which match locations will be preserved by a given expansion. Note that an expansion to a free variable can be pruned immediately, as any resulting abstraction will not be well-formed.

When expanding to an abstraction variable  $\alpha$ , if  $\alpha$  is a new variable not already present in the partial abstraction, the set of match locations is unchanged. If  $\alpha$  is an existing abstraction variable then this is a situation where the same variable is being used in more than one place, as in the *square* abstraction  $(\lambda \alpha. * \alpha \alpha)$ . In this case we restrict the match locations to the subset of locations where within a location all instances of  $\alpha$  match against the same subtree. Additionally, if a maximum arity is provided then an expansion that causes the abstraction to exceed this limit is not considered.

When considering each possible expanded abstraction  $A'_{??}$ , there is room for strong corpus-guidance. First, we don't need to consider any expansions that would result in zero match locations since all abstractions in this branch of search will have zero rewrite locations per Lemma 2 (lines 11-13). Furthermore, we use  $\bar{U}_{\mathcal{P}, \mathcal{R}}(A'_{??})$  to upper bound the utility achievable in this branch of search and discard it if it is less than our best utility so far (lines 14-16). Since each  $A'_{??}$  covers a disjoint set of match locations (except in the case of expanding into an abstraction variable), the set of match locations often drops rapidly and can allow for the calculation of a tight upper bound depending on the utility function. As a final step of pruning, if we can identify that  $A'_{??}$  is *strictly dominated* by some other abstraction  $A''_{??}$ , we may discard  $A'_{??}$  (lines 17-19). Note additionally that arity-zero abstractions were precomputed, then abstractions that match at a single location are safe to prune as well (as long as they don't have free variables) as arity-zero abstractions are always superior for single match locations.

For any partial abstraction  $A'_{??}$  that has not been pruned, we then check whether it is a complete abstraction (there are no remaining holes) or whether it is still a partial abstraction. Partial

abstractions get pushed to the priority queue ranked by their utility upper bound, and complete abstractions are used to update  $A_{best}$  and  $U_{best}$  if they have a higher utility than any prior complete abstractions.

Once there are no more partial abstractions remaining in the priority queue, the algorithm terminates. Note that since the algorithm maintains a best abstraction so-far, it can also be terminated early, making it an *any-time algorithm*. We also note that this algorithm is amenable to parallelization as it can be easily parallelized over the while loop on lines 4-29, especially since the algorithm remains sound even when the upper bound  $U_{best}$  used pruning is not always up to date.

This algorithm can be iterated to build up a library of abstractions.

Finally note that in our implementation, we employ structural hashing to simplify equality checks between subtrees and avoid re-doing work on multiple identical copies of a subtree. Our complete implementation of CTS for compression is provided at <https://github.com/mlb2251/stitch>.

---

**Algorithm 1** Corpus-guided top-down abstraction synthesis. Color-coded: **Upper bound pruning**, **Zero-usage pruning**, **Strict dominance pruning**

---

**Input:** Corpus of input programs  $\mathcal{P}$ , utility function  $U_{\mathcal{P}, \mathcal{R}}(A)$ , and utility upper bound function  $\bar{U}_{\mathcal{P}, \mathcal{R}}(A_{??})$

**Output:** The maximally compressive abstraction  $A_{best}$

```

1:  $Q \leftarrow$  Priority-Queue { ?? }           ▷ New priority queue with the single partial abstraction ???
2:  $A_{best} \leftarrow$  BEST-ARITY-ZERO-ABSTRACTION( $\mathcal{P}$ )           ▷ Initialize best abstraction so far
3:  $U_{best} \leftarrow U_{\mathcal{P}, \mathcal{R}}(A_{best})$ 
4: while non-empty( $Q$ ) do
5:    $A_{??} \leftarrow$  pop-max( $Q$ )           ▷ Next partial abstraction to expand
6:   if  $\bar{U}_{\mathcal{P}, \mathcal{R}}(A_{??}) \leq U_{best}$  then           ▷ Upper bound pruning
7:     continue
8:   end if
9:    $h \leftarrow$  CHOOSE-HOLE( $A_{??}$ )           ▷ Choose a hole to expand
10:  for  $(A'_{??}, M') \in$  EXPANSIONS( $A_{??}, h, \mathcal{P}$ ) do    ▷ abstraction  $A'_{??}$  and match locations  $M'$ 
11:    if length( $M'$ ) == 0 then           ▷ No match locations in corpus
12:      continue
13:    end if
14:    if  $\bar{U}_{\mathcal{P}, \mathcal{R}}(A'_{??}) \leq U_{best}$  then           ▷ Upper bound pruning
15:      continue
16:    end if
17:    if strictly-dominated( $A'_{??}, \mathcal{P}$ ) then           ▷ Strict dominance pruning
18:      continue
19:    end if
20:    if has-holes( $A_e$ ) then
21:       $Q \leftarrow Q \cup A'_{??}$            ▷ add partial abstraction to heap
22:    else
23:      if  $U_{\mathcal{P}, \mathcal{R}}(A'_{??}) > U_{best}$  then
24:         $U_{best} \leftarrow U_{\mathcal{P}, \mathcal{R}}(A'_{??})$ 
25:         $A_{best} \leftarrow A'_{??}$            ▷ new best complete abstraction
26:      end if
27:    end if
28:  end for
29: end while

```

---

## B PROOF OF CORRECTNESS OF LAMBDAUNIFY

As introduced in Section 3,  $\text{LAMBDAUNIFY}(A, e)$  returns a mapping from abstraction variables and holes to expressions  $[\alpha_i \rightarrow e'_i, \dots, ??_j \rightarrow u'_j, \dots]$  such that

$$(\lambda \alpha_i. \dots \lambda ??_j. \dots A) e'_i \dots u'_j \dots = e \quad (19)$$

through beta reduction. The  $\text{LAMBDAUNIFY}$  procedure is given in Fig. 4 and the definition of beta reduction with a modified upshifting operator is given in Fig. 5.

We can write Eq. (19) in terms of substitution as

$$[\alpha_i \rightarrow e'_i, \dots, ??_j \rightarrow u'_j, \dots] \circ A = e \quad (20)$$

Which can in turn be written in terms of  $\text{LAMBDAUNIFY}$  as

$$\text{LAMBDAUNIFY}(A, e) \circ A = e \quad (21)$$

To prove the correctness of  $\text{LAMBDAUNIFY}$ , we must show that Eq. (21) holds for all  $A$  and  $e$ . We do this in two ways: a written proof (below) of our correctness theorem (Theorem B.1), and a machine-verifiable proof in Coq. The Coq proof is given in `stitch.v` in the supplementary material and was proven with CoqIDE 8.15.2. The meanings of the definitions and lemmas in the Coq proof are commented. The Coq proof differs from the written proof in places because encoding the proof in Coq required coming up with some representations that worked well for Coq, though these differences are limited. For example, the Coq proof requires fully formalizing the behavior of MERGE and proofs related to it which is more simply explained in the written proof. When lemmas in the written proof below correspond directly to lemmas in the Coq proof, the corresponding Coq lemma is indicated in the written proof below.

### B.1 Well-formedness

We formalize a notion of *well-formedness* to capture the set of expressions that are accessible starting from an expression with no  $\&i$  variables then applying any series of upshifts and downshifts. We define  $\text{WF}_d(e)$ , the well-formedness of an expression  $e$  at a depth  $d$ , as follows:

$$\begin{aligned} \text{WF}_d \lambda.b &= \text{WF}_{d+1} b \\ \text{WF}_d (fx) &= \text{WF}_d f \wedge \text{WF}_d x \\ \text{WF}_d \$i &= \top \\ \text{WF}_d \&i &= \begin{cases} \top, & \text{if } i < d \\ \perp, & \text{if } i \geq d \end{cases} \\ \text{WF}_d t &= \top, \text{ for } t \in \mathcal{G}_{\text{sym}} \end{aligned}$$

Note that trivially any expression  $e$  without any  $\&i$  variables is well formed, as  $\&i$  variables are the only source of  $\perp$  in this definition.

If every expression in a mapping  $l$  is well-formed, we say that the mapping itself is well-formed, written ( $\text{WFMMap } l$ ).

### B.2 Relevant Lemmas

**The following two lemmas show that upshifting and downshifting preserve well-formedness.** In Coq these lemmas are `upshift_wf` and `downshift_wf`.

LEMMA 3.  $\forall e. \forall d. WF_d e \implies WF_d \uparrow_d e$

PROOF. We proceed by induction on  $e$ .

**Case**  $e = \lambda. b$ . Our goal is

$$WF_d \uparrow_d \lambda.b$$

We can move  $\uparrow_d$  and  $WF_d$  into the lambda by unfolding their definitions to get

$$WF_{d+1} \uparrow_{d+1} b$$

We also have the premise  $WF_d \lambda.b$ , where we can likewise move the  $WF_d$  into the lambda to get  $WF_{d+1} b$ . The inductive hypothesis  $\forall d. WF_d b \implies WF_d \uparrow_d b$  can be instantiated with  $d$  as  $d + 1$  to get  $WF_{d+1} b \implies WF_{d+1} \uparrow_{d+1} b$ , the conclusion of which is precisely our goal and the premise of which is precisely our other premise.

**Case**  $e = f x$ . Our goal is

$$WF_d \uparrow_d (f x)$$

We can move  $\uparrow_d$  and  $WF_d$  into the application by unfolding their definitions to get

$$(WF_d \uparrow_d f) \wedge (WF_d \uparrow_d x)$$

We can use our inductive hypotheses  $WF_d f \implies WF_d \uparrow_d f$  to prove the left side of this conjunction if we can prove  $WF_d f$ . We have the premise that  $WF_d (f x)$  from this overall induction case, which by definition of  $WF_d$  can only be true if  $WF_d f \wedge WF_d x$ . Thus we can prove the left side of the conjunction, and the proof of the right side is identical with  $f$  replaced by  $x$  using the other inductive hypothesis.

**Case**  $e = \$i$ . Our goal is

$$WF_d \uparrow_d \$i$$

We will handle this in 2 cases:

- Case  $i < d$ .

$$WF_d \uparrow_d \$i$$

$$= WF_d \$i$$

This is true because  $\$i$  indices are always well-formed.

- Case  $i \geq d$ .

$$WF_d \uparrow_d \$i$$

$$= WF_d \$(i + 1)$$

This is true because  $\$i$  indices are always well-formed.

**Case**  $e = \&i$ . Our goal is

$$WF_d \uparrow_d \&i$$

We will handle our goal in 2 cases:

- Case  $i + 1 \neq d$ .

$$\begin{aligned} & \text{WF}_d \uparrow_d \$i \\ &= \text{WF}_d \&(i+1) \end{aligned}$$

We know from our premise that  $\text{WF}_d \&i$  which means  $i < d$ , so  $i + 1 \leq d$ . Combining this with our knowledge from the case branch that  $i + 1 \neq d$ , we know that  $i + 1 < d$  which means that  $\&(i + 1)$  is well formed.

- Case  $i + 1 = d$ .

$$\begin{aligned} & \text{WF}_d \uparrow_d \&i \\ &= \text{WF}_d \$i \end{aligned}$$

This is true because  $\$i$  indices are always well-formed.

**Case  $e = t$** . Our goal is  $\text{WF}_d \uparrow_d t$  which simplifies by definition of  $\uparrow_d$  to  $\text{WF}_d t$  which is true by definition of  $\text{WF}_d$ .

□

LEMMA 4.  $\forall e. \forall d. \text{WF}_d e \implies \text{WF}_d \downarrow_d e$

PROOF. We proceed by induction on  $e$ . The cases  $e = t$ ,  $e = (f x)$  and  $e = \lambda.b$  are identical to those in Lemma 3 but with  $\uparrow$  replaced with  $\downarrow$  so they will be omitted.

**Case  $e = \$i$** . Our goal is

$$\text{WF}_d \downarrow_d \$i$$

We will handle this in 3 cases:

- Case  $i < d$ .

$$\begin{aligned} & \text{WF}_d \downarrow_d \$i \\ &= \text{WF}_d \$i \end{aligned}$$

This is true because  $\$i$  indices are always well-formed.

- Case  $i > d$ .

$$\begin{aligned} & \text{WF}_d \downarrow_d \$i \\ &= \text{WF}_d \$i \end{aligned}$$

This is true because  $\$i$  indices are always well-formed.

- Case  $i = d$ .

$$\begin{aligned} & \text{WF}_d \downarrow_d \$i \\ &= \text{WF}_d \&(i-1) \end{aligned}$$

Since  $i = d$  we know that  $i - 1 < d$  which means  $\&(i - 1)$  is well-formed.

**Case  $e = \&i$** . Our goal is

$$\text{WF}_d \downarrow_d \&i$$

which simplifies to

$$\text{WF}_d \&(i-1)$$

We know from our premise that  $\text{WF}_d \&i$  so  $i < d$  which means  $i - 1 < d$  so  $\&(i - 1)$  is well-formed.

□

**The following lemma shows that upshifting is an inverse of downshifting.** In Coq this lemma is `upshift_downshift`.

LEMMA 5.  $\forall e. \forall d. WF_d e \implies \uparrow_d \downarrow_d e = e$

PROOF. We proceed by induction on  $e$ .

**Case  $e = \lambda. b$ .** Our goal is

$$\uparrow_d \downarrow_d \lambda. b = \lambda. b$$

By the definitions of  $\uparrow_d$  and  $\downarrow_d$  we can propagate them into the body of a lambda while incrementing  $d$ :

$$\lambda. \uparrow_{d+1} \downarrow_{d+1} b = \lambda. b$$

We also have the additional premise  $WF_d \lambda. b$  which simplifies to  $WF_{d+1} b$ .

Since our inductive hypothesis  $\forall d. WF_d b \implies \uparrow_d \downarrow_d b = b$  is universally quantified over  $d$  we can instantiate it with  $d + 1$  in place of  $d$  and discharge its precondition with our other premise  $WF_{d+1} b$  to get  $\uparrow_{d+1} \downarrow_{d+1} b = b$  and apply this to our goal to get

$$\lambda. b = \lambda. b$$

which is trivially true.

**Case  $e = f x$ .** Our goal is

$$\uparrow_d \downarrow_d (f x) = f x$$

By the definitions of  $\uparrow_d$  and  $\downarrow_d$  we can propagate them into both sides of the application:

$$(\uparrow_d \downarrow_d f)(\uparrow_d \downarrow_d x) = f x$$

We also have the additional premise  $WF_d (f x)$  can be broken into the two premises  $WF_d f$  and  $WF_d x$ .

We can then apply our inductive hypotheses  $\forall d. WF_d f \implies \uparrow_d \downarrow_d f = f$  and  $\forall d. WF_d x \implies \uparrow_d \downarrow_d x = x$  with these premises to rewrite our goal to

$$f x = f x$$

which is trivially true.

**Case  $e = \$i$ .** Our goal is

$$\uparrow_d \downarrow_d \$i = \$i$$

We will handle this in 3 cases:

- *Case  $i < d$ .*

$$\uparrow_d \downarrow_d \$i = \$i$$

Unfold definition of  $\downarrow_d$

$$\uparrow_d \$i = \$i$$

Unfold definition of  $\uparrow_d$

$$\$i = \$i$$

- Case  $i > d$ .

$$\uparrow_d \downarrow_d \$i = \$i$$

Unfold definition of  $\downarrow_d$

$$\uparrow_d \$i = \$i$$

Unfold definition of  $\uparrow_d$  noting that  $(i - 1) \geq d$

$$\$( (i - 1) + 1 ) = \$i$$

Simplify

$$\$i = \$i$$

- Case  $i = d$ .

$$\uparrow_d \downarrow_d \$i = \$i$$

Unfold definition of  $\downarrow_d$

$$\uparrow_d \& (i - 1) = \$i$$

Unfold definition of  $\uparrow_d$  noting that  $(i - 1) + 1 = d$

$$\$( (i - 1) + 1 ) = \$i$$

Simplify

$$\$i = \$i$$

**Case  $e = \&i$ .** Our goal is

$$\uparrow_d \downarrow_d \&i = \&i$$

Unfolding the definition of  $\downarrow_d$  yields

$$\uparrow_d \& (i - 1) = \&i$$

Since we know that  $\text{WF}_d \&i$  by our premise, we know that  $i < d$  so  $(i - 1) + 1 < d$  so  $(i - 1) + 1 \neq d$  so  $\uparrow_d$  unfolds to:

$$\& ( (i - 1) + 1 ) = \&i$$

which simplifies to

$$\&i = \&i$$

which is trivially true.

**Case  $e = t$ .** Our goal is

$$\uparrow_d \downarrow_d t = t$$

Neither  $\uparrow_d$  nor  $\downarrow_d$  have any effect on a grammar primitive  $t$ , so this is trivially true.  $\square$

**The following lemma shows that any result of `LAMBDAUNIFY` is a well-formed mapping.** In Coq this lemma is called `lu_wf`.

LEMMA 6.  $\forall A. \forall e. \text{WFMap } \text{LAMBDAUNIFY}(A, e)$

**PROOF.** This is a straightforward proof by induction. In the  $\alpha_i$  and  $\text{??}_i$  cases LAMBDAUNIFY adds expressions from the original expression to the mapping, which are well-formed as they can't contain  $\&i$  variables. In the lambda case DOWNSHIFTALL shifts variables using  $\downarrow$  which is guaranteed to preserve well-formedness by Lemma 4. In the application case MERGE simply combines mappings from the function and argument calls to LAMBDAUNIFY which are well-formed by induction, resulting in an overall well-formed mapping.  $\square$

**The following lemma follows from Lemma 5 and shows that UPSHIFTALL is an inverse of DOWNSHIFTALL.** In Coq this lemma is `upshift_all_downshift_all`.

LEMMA 7.  $\text{WFMap } l \implies \text{UPSHIFTALL}(\text{DOWNSHIFTALL}(l)) = l$

**PROOF.** Writing  $l$  as  $[\alpha_i \rightarrow e'_i, \text{??}_j \rightarrow u'_j, \dots]$  and unfolding the definitions of UPSHIFTALL DOWNSHIFTALL we get:

$$[\alpha_i \rightarrow \uparrow_0 \downarrow_0 e'_i, \text{??}_j \rightarrow \uparrow_0 \downarrow_0 u'_j, \dots] = [\alpha_i \rightarrow e'_i, \text{??}_j \rightarrow u'_j, \dots]$$

This equality will hold if  $\uparrow_0 \downarrow_0 u'_j = u'_j$  and  $\uparrow_0 \downarrow_0 e'_i = e'_i$  for all  $u'_j$  and  $e'_i$ .  $\text{WFMap } l$  tells us that all expressions  $u'_j$  and  $e'_i$  are well-formed and thus satisfy the precondition of Lemma 5, which proves these equalities true.  $\square$

**The following two lemmas allow us to "forget" one side of a MERGE when it won't have any effect on the substitution.** In Coq the first lemma is `merge_forget3` and the second is captured by the definition `Compatible`.

LEMMA 8. *If  $l_1$  is a mapping that binds every abstraction variable and hole in  $A$ , then  $\text{MERGE}(l_1, l_2) \circ A = l_1 \circ A$ .*

**PROOF.**  $l_1$  already assigns to every abstraction variable and hole, and merging in  $l_2$  can't overwrite these assignments since MERGE fails when the same variable or hole is assigned to two different expressions. Thus merging in  $l_2$  has no effect on the resulting expression from substitution.  $\square$

LEMMA 9. *If  $l_2$  is a mapping that binds every abstraction variable and hole in  $A$ , then  $\text{MERGE}(l_1, l_2) \circ A = l_2 \circ A$ .*

**PROOF.** This proof is identical in structure to Lemma 8.  $l_2$  already assigns to every abstraction variable and hole, and merging in  $l_1$  can't overwrite these assignments since MERGE fails when the same variable or hole is assigned to two different expressions. Thus merging in  $l_1$  has no effect on the resulting expression from substitution.  $\square$

**The following lemma shows that any result of LAMBDAUNIFY satisfies the precondition of Lemma 9 and Lemma 8.**

LEMMA 10. *LAMBDAUNIFY( $A, e$ ) returns a mapping that binds all abstraction variables  $\alpha_i$  and all holes  $\text{??}_i$  in  $A$ .*

**PROOF.** This is a straightforward proof by induction. In the  $\alpha_i$  and  $\text{??}_i$  cases LAMBDAUNIFY binds the variable and hole respectively, in the lambda case DOWNSHIFTALL has no effect on which abstraction variables are bound (and it binds all the variables in the body by induction), and in the application case MERGE results in combining the function and argument mappings (which bind all variables by induction) to result in a mapping which binds all variables that appear in both the function and argument.  $\square$

### B.3 Correctness Theorem

In Coq this theorem is called correctness.

THEOREM B.1.  $\forall A. \forall e. \text{LAMBDAUNIFY}(A, e) \circ A = e$

PROOF. We proceed by induction over the proof tree of  $\text{LAMBDAUNIFY}(A, e)$ .

*Case U-ABSVAR.* Our goal is  $\text{LAMBDAUNIFY}(A, \alpha) \circ \alpha = e$  which simplifies by definition of  $\text{LAMBDAUNIFY}$  to  $[\alpha \rightarrow e] \circ \alpha = e$  which is true by the definition of substitution.

*Case U-HOLE.* Our goal is  $\text{LAMBDAUNIFY}(A, ??) \circ ?? = e$  which simplifies by definition of  $\text{LAMBDAUNIFY}$  to  $[?? \rightarrow e] \circ ?? = e$  which is true by the definition of substitution.

*Case U-APP.* Our goal is

$$\text{LAMBDAUNIFY}(A_1 A_2, e_1 e_2) \circ (A_1 A_2) = e_1 e_2$$

This simplifies by definition of  $\text{LAMBDAUNIFY}$  to

$$\text{MERGE}(\text{LAMBDAUNIFY}(A_1, e_1), \text{LAMBDAUNIFY}(A_2, e_2)) \circ (A_1 A_2) = e_1 e_2$$

We can distribute the substitution over the application (by definition of substitution in Fig. 5 (Left)) to get

$$\begin{aligned} & ((\text{MERGE}(\text{LAMBDAUNIFY}(A_1, e_1), \text{LAMBDAUNIFY}(A_2, e_2)) \circ A_1) \\ & (\text{MERGE}(\text{LAMBDAUNIFY}(A_1, e_1), \text{LAMBDAUNIFY}(A_2, e_2)) \circ A_2)) \\ & = e_1 e_2 \end{aligned}$$

We can then forget the irrelevant part of each merge using Lemma 9 and Lemma 8 where the precondition (binding all abstraction variables and holes) is satisfied by Lemma 10:

$$\begin{aligned} & ((\text{LAMBDAUNIFY}(A_1, e_1) \circ A_1) \\ & (\text{LAMBDAUNIFY}(A_2, e_2) \circ A_2)) \\ & = e_1 e_2 \end{aligned}$$

We can then finish the proof by directly applying our inductive hypotheses  $\text{LAMBDAUNIFY}(A_1, e_1) \circ A_1 = e_1$  and  $\text{LAMBDAUNIFY}(A_2, e_2) \circ A_2 = e_2$ .

*Case U-LAM.* Our goal is

$$\text{LAMBDAUNIFY}(\lambda. A, \lambda. e) \circ \lambda. A = \lambda. e$$

This simplifies by definition of  $\text{LAMBDAUNIFY}$  to

$$\text{DOWNSHIFTALL}(\text{LAMBDAUNIFY}(A, e)) \circ \lambda. A = \lambda. e$$

We can move the substitution under the lambda while inserting  $\text{UPSHIFTALL}$  as per the definition of substitution

$$\lambda. \text{UPSHIFTALL}(\text{DOWNSHIFTALL}(\text{LAMBDAUNIFY}(A, e))) \circ A = \lambda. e$$

Since we know that any result of LAMBDAUNIFY is well-formed by Lemma 6 we can apply Lemma 7 to get

$$\lambda. \text{LAMBDAUNIFY}(A, e) \circ A = \lambda. e$$

We can then apply our inductive hypothesis  $\text{LAMBDAUNIFY}(A, e) \circ A = e$  to get

$$\lambda. e = \lambda. e$$

and we are done.

*Case U-SAME.* Our goal is  $\text{LAMBDAUNIFY}(e, e) \circ e = e$  which simplifies by definition of LAMBDAUNIFY to  $[] \circ e = e$  which is trivially true as the empty substitution has no effect.  $\square$

Received 2022-07-07; accepted 2022-11-07