TBar: Revisiting Template-based Automated Program Repair

Kui Liu, Anil Koyuncu, Dongsun Kim, Tegawendé F. Bissyandé
University of Luxembourg, Luxembourg
{kui.liu,anil.koyuncu,dongsun.kim,tegawende.bissyande}@uni.lu

ABSTRACT
We revisit the performance of template-based APR to build comprehensive knowledge about the effectiveness of fix patterns, and to highlight the importance of complementary steps such as fault localization or donor code retrieval. To that end, we first investigate the literature to collect, summarize and label recurrently-used fix patterns. Based on the investigation, we build TBar, a straightforward APR tool that systematically attempts to apply these fix patterns to program bugs. We thoroughly evaluate TBar on the Defects4J benchmark. In particular, we assess the actual qualitative and quantitative diversity of fix patterns, as well as their effectiveness in yielding plausible or correct patches. Eventually, we find that, assuming a perfect fault localization, TBar correctly/plausibly fixes 74/101 bugs. Replicating a standard and practical pipeline of APR assessment, we demonstrate that TBar correctly fixes 43 bugs from Defects4J, an unprecedented performance in the literature (including all approaches, i.e., template-based, stochastic mutation-based or synthesis-based APR).

CCS CONCEPTS
• Software and its engineering → Software verification and validation; Software defect analysis; Software testing and debugging.

KEYWORDS
Automated program repair, fix pattern, empirical assessment.

ACM Reference Format:

1 INTRODUCTION
Automated Program Repair (APR) has progressively become an essential research field. APR research is indeed promising to improve modern software development by reducing the time and costs associated with program debugging tasks. In particular, given that faults in software cause substantial financial losses to the software industry [8, 54], there is a momentum in minimizing the time-to-fix intervals by APR. Recently, various APR approaches [10, 11, 17, 18, 21, 23, 25, 26, 29, 31, 36, 38, 39, 41, 42, 44, 51, 53, 67, 69, 76, 77] have been proposed, aiming at reducing manual debugging efforts through automatically generating patches.

An early strategy of APR is to generate concrete patches based on fix patterns [23] (also referred to as fix templates [40] or program transformation schemas [17]). This strategy is now common in the literature and has been implemented in several APR systems [13, 17, 23, 24, 38–40, 50, 62]. Kim et al. [23] showed the usefulness of fix patterns with PAR. Saha et al. [62] later proposed ELIXIR by adding three new patterns on top of PAR [23]. Durieux et al. [13] proposed NPEfix to repair null pointer exception bugs, using nine pre-defined fix patterns. Long et al. designed Genesis [41] to infer fix patterns for specific three classes of defects. Liu and Zhong [40] explored posts from Stack Overflow to mine fix patterns for APR. Hua et al. proposed SketchFix [17], a runtime on-demand APR tool with six pre-defined fix patterns. Recently, Liu et al. [39] used the fix patterns of FindBugs static violations [35] to fix semantic bugs. Concurrently, Ghanbari and Zhang [15] showed that straightforward application of fix patterns (i.e., mutators) on Java bytecode is effective for repair. They do not, however, provide a comprehensive assessment of the repair performance yielded by each implemented mutator.

Although the literature has reported promising results with fix patterns-based APR, to the best of our knowledge, no extensive assessment on the effectiveness of various patterns is performed. A few most recent approaches [17, 39, 40] reported which benchmark bugs are fixed by each of their patterns. Nevertheless, many relevant questions on the effectiveness of fix patterns remain unanswered.

This paper. Our work thoroughly investigates to what extent fix patterns are effective for program repair. In particular, emphasizing on the recurrence of some patterns in APR, we dissect their actual contribution to repair performance. Eventually, we explore three aspects of fix patterns:

• Diversity: How diverse are the fix patterns used by the state-of-the-art? We survey the literature to identify and summarize the available patterns with a clear taxonomy.
• Repair performance: How effective are the different patterns? In particular, we investigate the variety of real-world bugs that can be fixed, the dissection of repair results, and their tendency to yield plausible or correct patches.
• Sensitivity to fault localization noise: Are all fix patterns similarly sensitive to the false positives yielded by fault localization tools? We investigate sensitivity by assessing plausible patches as well as the suspiciousness rank of correctly-fixed bug locations.

Towards realizing this study, we implement an automated patch generation system, TBar (Template-Based automated program repair), with a super-set of fix patterns that are collected, summarized, curated and labeled from the literature data. We evaluate TBar on the Defects4J [20] benchmark, and provide the replication package in a public repository: https://github.com/SerVal-DTF/TBar.

Overall, our investigations have yielded the following findings:

Permission to make digital or hard copies of all or part 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 components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ISSSTA ’19, July 15–19, 2019, Beijing, China
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6224-5/19/07... $15.00
https://doi.org/10.1145/3293882.3330577
(1) **Record performance**: TBar creates a new higher baseline of repair performance: 74/101 bugs are correctly/plausibly fixed with perfect fault localization information and 43/81 bugs are fixed with realistic fault localization output, respectively.

(2) **Fix pattern selection**: Most bugs are correctly fixed only by a single fix pattern while other patterns generate plausible patches. This implies that appropriate pattern prioritization can prevent from plausible/incorrect patches. Otherwise, APR tools might be overfitted in plausible but incorrect patches.

(3) **Fix ingredient retrieval**: It is challenging for template-based APR to select appropriate donor code, which is an ingredient of patch generation when using fix patterns. Inappropriate donor code may cause plausible but incorrect patch generation. This motivates a new research direction: donor code prioritization.

(4) **Fault localization noise**: It turns out that fault localization accuracy has a large impact on repair performance when using fix patterns in APR (e.g., applying a fix pattern to incorrect location yields plausible/incorrect patches).

### 2. FIX PATTERNS

For this study, we systematically review\(^1\) the APR literature to identify approaches that leverage fix patterns. Concretely, we consider the program repair website [3], a bibliography survey of APR [52], proceedings of software engineering conference venues and journals as the source of relevant literature. We focus on approaches dealing with Java program bugs, and manually collect, from the paper descriptions as well as the associated artifacts, all pattern instances that are explicitly mentioned. Table 1 summarizes the identified relevant literature and the quantity of identified fix patterns targeting Java programs. Note that the techniques described in the last four papers (i.e., HDRepair, ssFix, CapGen, and SimFix papers) do not directly use fix patterns: they leverage code change operators or rules, which we consider similar to using fix patterns.

**Table 1: Literature review on fix patterns for Java programs.**

<table>
<thead>
<tr>
<th>Authors</th>
<th>APR tool name</th>
<th># of fix patterns</th>
<th>Publication Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liu et al. [39]</td>
<td>AVATAR</td>
<td>13</td>
<td>SANER 2019</td>
</tr>
<tr>
<td>Pan et al. [55]</td>
<td>PAR</td>
<td>10 (16(^*))</td>
<td>ECSE 2009</td>
</tr>
<tr>
<td>Kim et al. [23]</td>
<td>PAR</td>
<td>3</td>
<td>ISST 2016</td>
</tr>
<tr>
<td>Martinez et al. [49]</td>
<td>PAR</td>
<td>9</td>
<td>ICSE 2013</td>
</tr>
<tr>
<td>Durieux et al. [13]</td>
<td>NPEfix</td>
<td>9</td>
<td>SANER 2017</td>
</tr>
<tr>
<td>Long et al. [41]</td>
<td>Genesis</td>
<td>3 (108(^*))</td>
<td>FSE 2017</td>
</tr>
<tr>
<td>D. Le et al. [25]</td>
<td>SY</td>
<td>4</td>
<td>FSE 2017</td>
</tr>
<tr>
<td>Saha et al. [62]</td>
<td>ELIXIR</td>
<td>8 (11(^*))</td>
<td>ASE 2017</td>
</tr>
<tr>
<td>Hua et al. [17]</td>
<td>SketchFix</td>
<td>6</td>
<td>ICSE 2018</td>
</tr>
<tr>
<td>Liu and Zhong [40]</td>
<td>ssFix</td>
<td>12</td>
<td>SANER 2018</td>
</tr>
<tr>
<td>Koyuncu et al. [24]</td>
<td>FixNiner</td>
<td>28</td>
<td>UC Tech Report 2018</td>
</tr>
<tr>
<td>Liu et al. [35]</td>
<td>-</td>
<td>174</td>
<td>TSE 2018</td>
</tr>
<tr>
<td>Rolim et al. [60]</td>
<td>REVISAR</td>
<td>9</td>
<td>UFESR Tech Report 2018</td>
</tr>
<tr>
<td>Liu et al. [29]</td>
<td>HDRepair(^\d)</td>
<td>11</td>
<td>SANER 2016</td>
</tr>
<tr>
<td>Xin and Reiss [44]</td>
<td>ssFix(^\d)</td>
<td>34</td>
<td>ASE 2017</td>
</tr>
<tr>
<td>Wen et al. [40]</td>
<td>CapGen(_\d)</td>
<td>30</td>
<td>ICSE 2018</td>
</tr>
<tr>
<td>Jiang et al. [18]</td>
<td>SimFix(_\d)</td>
<td>16</td>
<td>ISST 2018</td>
</tr>
</tbody>
</table>

\(^{\text{1}}\): In the PAR paper [29], 10 fix patterns are presented, but 16 fix patterns are released online [3]. In Genesis, 108 code transformation schemas are inferred for three kinds of defects. In ELIXIR, there is one fix pattern that consists of four sub-fix patterns.

#### 2.1 Fix Patterns Inference

Fix patterns have been explored with the following four ways:

(1) **Manual Summarization**: Pan et al. [55] identified 27 fix patterns from patches of five Java projects to characterize the fix ingredients of patches. They do not however apply the identified patterns to fix actual bugs. Motivated by this work, Kim et al. [25] summarized 10 fix patterns manually extracted from 62,656 human-written patches collected from Eclipse JDT.

(2) **Mining**: Long et al. [41] proposed Genesis, to infer fix patterns for three kinds of defects from existing patches. Liu and Zhong [40] explored fix patterns from Q&A posts in Stack Overflow. Koyuncu et al. [24] mined fix patterns at the AST level from patches by using code change differentiating tool [14]. Liu et al. [35] and Rolim et al. [60] proposed to mine fix patterns for static analysis violations. In general, mining approaches yield a large number of fix patterns, which are not always about addressing deviations in program behavior. For example, many patterns are about code style [39]. Recently, with AVATAR [39], we proposed an APR tool that considers static analysis violation fix patterns to fix semantic bugs.

(3) **Pre-definition**: Durieux et al. [13] pre-defined 9 repair actions for null pointer exceptions by unifying the related fix patterns proposed in previous studies [12, 22, 45]. On the top of PAR [23], Saha et al. [62] further defined 3 new fix patterns to improve the repair performance. Hua et al. [17] proposed an APR tool with six pre-defined so-called code transformation schemas. We also consider operator mutations [49] as pre-defined fix patterns, as the number of operators and mutation possibilities is limited and pre-set. Xin and Reiss [74] proposed an approach to fixing bugs with 34 predefined code change rules at the AST level. Ten of the rules are not for transforming the buggy code but for the simple replacement of multi-statement code fragments. We discard these rules from our study to limit bias.

(4) **Statistics**: Besides formatted fix patterns, researchers [18, 69] also explored to automate program repair with code change instructions (at the abstract syntax tree level) that are statistically recurrent in existing patches [18, 37, 48, 68, 81]. The strategy is then to select the top-n most frequent code change instructions as fix ingredients to synthesize patches.

#### 2.2 Fix Patterns Taxonomy

After manually assessing all fix patterns presented in the literature (cf. Table 1), we identified 15 categories of patterns labeled based on the code context (e.g., a cast expression), the code change actions (e.g., insert an “if” statement with “instanceof” check) as well as the targets (e.g., ensure the program will no throw a `ClassCastException`). A given category may include one or several specialized sub-categories. Below, we present the labeled categories and provide the associated 35 **Code Change Patterns** described in simplified GNU diff pattern for easy understanding.

**FP1: Insert Cast Checker**: Inserting an `instanceof` check before one buggy statement if this statement contains at least one unchecked cast expression. **Implemented in**: PAR, Genesis, AVATAR, SOFix\(^\text{\dagger}\), HDRepair\(^\text{\dagger}\), SketchFix\(^\text{\dagger}\), CapGen\(^\text{\dagger}\), and SimFix\(^\text{\dagger}\).

\[\text{if } (\text{exp } \text{instanceof } \text{T}) \{ \text{var } = (\text{T}) \text{exp}; \ldots \} \]

where `exp` is an expression (e.g., a variable expression) and `T` is the casting type, while “\ldots” means the subsequent statements dependent on the variable `var`. Note that, “\dagger” denotes that the fix pattern is not specifically illustrated in the corresponding APR tools
Since the tools have some abstract fix patterns that can cover the fix pattern. The same notation applies to the following descriptions.

**FP2. Insert Null Pointer Checker.** Inserting a null check before a buggy statement if, in this statement, a field or an expression (of non-primitive data type) is accessed without a null pointer check. Implemented in: PAR, ELIXIR, NPEfix, Genesis, FixMiner, AVATAR, HDRepair\(^7\), SOFix\(^1\), SketchFix\(^1\), CapGen\(^1\), and SimFix\(^1\).

\[
FP2.1:\quad + \text{if (exp} \neq \text{null)} \{ \\
\quad \quad \quad \text{...exp...} \quad \text{...} \\
\quad + \}
\]
\[
FP2.2:\quad + \text{if (exp} = \text{null)} \text{return DEFAULT VALUE;} \\
\quad \text{...exp...}
\]
\[
FP2.3:\quad + \text{if (exp} = \text{null)} \text{exp} = \text{exp1;} \\
\quad \text{...exp...}
\]
\[
FP2.4:\quad + \text{if (exp} = \text{null)} \text{continue;} \\
\quad \text{...exp...}
\]
\[
FP2.5:\quad + \text{if (exp} = \text{null)} \\
\quad \quad \quad \text{throw new IllegalArgumentException(...);} \\
\quad \text{...exp...}
\]

where DEFAULT VALUE is set based on the return type (RT) of the encompassing method as below:

\[
\text{DEFAULT VALUE} = \begin{cases} 
\text{false}, & \text{if RT is boolean;} \\
0, & \text{if RT is primitive type;} \\
\text{new String()}, & \text{if RT is String;} \\
\text{return}, & \text{if RT is void.} \\
\text{null, otherwise.} 
\end{cases}
\]

exp1 is a compatible expression in the buggy program (i.e., that has the same data type as exp). FP2.4 is specific to the case of a buggy statement within a loop (i.e., for or while).

**FP3. Insert Range Checker.** Inserting a range checker for the access of an array or collection if it is unchecked. Implemented in: PAR, ELIXIR, Genesis, SketchFix, AVATAR, SOFix\(^1\) and SimFix\(^1\).

\[
\text{FP3.1:} \\
\quad \text{if (index} < \text{exp.length}) \{ \\
\quad \quad \quad \text{...exp[index]}... \quad \text{...} \\
\quad \quad \quad \text{...} \\
\quad + \}
\]

where exp is an expression representing an array or collection.

**FP4. Insert Missed Statement.** Inserting a missing statement before, or after, or surround a buggy statement. The statement is either an expression statement with a method invocation, or a return/try-catch/if statement. Implemented in: ELIXIR, HDRepair, SOFix, SketchFix, CapGen, FixMiner, and SimFix.

\[
\text{FP4.1:} \\
\quad \text{method(exp);}
\]
\[
\text{FP4.2:} \\
\quad \text{return DEFAULT VALUE;}
\]
\[
\text{FP4.3:} \\
\quad \text{try} \\
\quad \quad \quad \text{statement; ...} \\
\quad \quad \quad \text{...} \\
\quad \quad \quad \text{...} \\
\quad \quad \quad \text{...} \\
\quad + \]
\[
\text{FP4.4:} \\
\quad \text{if (conditional exp)} \\
\quad \quad \text{statement; ...} \\
\quad + \]

where exp is an expression from a buggy statement. It may be empty if the method does not take any argument. FP4.4 excludes three fix patterns (FP1, FP2, and FP3) that are used with specific contexts.

**FP5. Mutate Class Instance Creation.** Replacing a class instance creation expression with a cast super.clone() method invocation if the class instance creation is in an overridden clone method. Implemented in: AVATAR.

\[
\text{FP5.1:} \\
\quad \text{public Object clone()} \\
\quad \quad \text{...new T();} \\
\quad + \quad \text{(T) super.clone();}
\]

where T is the class name of the current class containing the buggy statement.

**FP6. Mutate Conditional Expression.** Mutating a conditional expression that returns a boolean value (i.e., true or false) by either updating it, or removing a sub conditional expression, or inserting a new conditional expression into it. Implemented in: PAR, ssFix, S3, HDRepair, ELIXIR, SketchFix, CapGen, SimFix, and AVATAR.

\[
\text{FP6.1:} \\
\quad \quad \quad \text{...condExp1...} \\
\quad + \quad \text{...condExp2...}
\]

where condExp1 and condExp2 are conditional expressions. Op is the logical operator ‘||’ or ‘&&’. The mutation of operators in conditional expressions is not summarized in this fix pattern but in FP11.

**FP7. Mutate Data Type.** Replacing the data type in a variable declaration or a cast expression with another data type. Implemented in: PAR, ELIXIR, FixMiner, SOFix, CapGen, SimFix, AVATAR, and HDRepair\(^7\).

\[
\text{FP7.1:} \\
\quad \quad \quad \text{T1 var ...;} \\
\quad + \quad \text{T2 var ...;}
\]
\[
\text{FP7.2:} \\
\quad \quad \quad \text{...(T1) exp...} \\
\quad + \quad \text{...(T2) exp...;}
\]

where both T1 and T2 denote two different data types. exp means the being casted expression (including variable).

**FP8. Mutate Integer Division Operation.** Mutating the integer division expressions to return a float value, by mutating its divisor or divider to make them be of type float. Released by Liu et al. [35], it is not implemented in any APR tool yet.

\[
\text{FP8.1:} \\
\quad \quad \quad \text{...dividend / divisor...} \\
\quad + \quad \text{...dividend / (double or float) divisor...}
\]
\[
\text{FP8.2:} \\
\quad \quad \quad \text{...dividend / divisor...} \\
\quad + \quad \text{...(double or float) dividend / divisor...}
\]
\[
\text{FP8.3:} \\
\quad \quad \quad \text{...dividend / divisor...} \\
\quad + \quad \text{...(1.0 / divisor) * dividend...}
\]

where dividend and divisor are integer number literals or integer-returned expressions (including variables).

**FP9. Mutate Literal Expression.** Mutating boolean, number, or String literals in a buggy statement with other relevant literals, or correspondingly-typed expressions. Implemented in: HDRepair, S3, FixMiner, SketchFix, CapGen, SimFix and ssFix\(^7\).

\[
\text{FP9.1:} \\
\quad \quad \quad \text{...literal1...} \\
\quad + \quad \text{...literal2...}
\]
\[
\text{FP9.2:} \\
\quad \quad \quad \text{...literal1...} \\
\quad + \quad \text{...exp...}
\]

where literal1 and literal2 are of the same type literals, but having different values (e.g., literal1 is true, literal2 is false). exp denotes any expression value of the same type as literal1.

**FP10. Mutate Method Invocation Expression.** Mutating the buggy method invocation expression by adapting its method name or arguments. This pattern consists of four sub fix patterns:

1. Replacing the method name with another one which has a compatible return type and parameter type(s) as the buggy method that was invoked.
2. Replacing at least one argument with another expression which has a compatible data type. Replacing a literal or variable is not included in this fix pattern, but rather in FP9 and FP13 respectively.
3. Removing argument(s) if the method invocation has the suitable overridden methods.
(4) Inserting argument(s) if the method invocation has the suitable overridden methods.

Implemented in: PAR, HDRepair, ssFix, ELIXIR, FixMiner, SOFix, SketchFix, CapGen, and SimFix.

FP10.1: - ....method1(args)....
+ ....method2(args)....
FP10.2: - ....method1(arg1, arg2, ...)....
+ ....method1(arg1, arg2, ...)....
+ ....method2(arg1, arg2, ...)....
FP10.3: - ....method1(arg1, arg2, ...)....
+ ....method1(arg1, arg2, ...)....
+ ....method2(arg1, arg2, ...)....

where method1 and method2 are the names of invoked methods. arg1, arg2 and arg3 denote the argument expressions in the method invocation. Note that, code changes on class instance creation, constructor and super constructor expressions are also included in these four fix patterns.

FP11. Mutate Operators. Mutating an operation expression by mutating its operator(s). We divide this fix pattern into three sub-fix patterns following the operator types and mutation actions.

(1) Replacing one operator with another operator from the same operator class (e.g., relational or arithmetic).

(2) Changing the priority of arithmetic operators.

(3) Replacing Instanceof operator with in/equality operators.

Implemented in: HDRepair, ssFix, ELIXIR, S3, jMutRepair, SOFix, SketchFix, CapGen, SimFix, AVATAR, and PAR.

FP11.1: - ....exp1 Op1 exp2....
+ ....exp1 Op2 exp2....
FP11.2: - ....exp1 Op1 (exp2 Op2 exp3)....
+ ....exp1 Op2 (exp2 Op2 exp3)....
FP11.3: - ....exp instanceof T....
+ ....exp != null....
+ ....exp = null....

where exp denotes the expressions in the operation and Op is the associated operator.


- return exp1;
- return exp2;

where exp1 and exp2 represent the returned expressions.

FP13. Mutate Variable. Replacing a variable in a buggy statement with a compatible expression (including variables and literals). Implemented in: S3, SOFix, FixMiner, SketchFix, CapGen, SimFix, AVATAR, and ssFix.

FP13.1: - ....var1....
+ ....var2....
+ ....exp....

where var1 denotes a variable in the buggy statement. var2 and exp represent respectively a compatible variable and expression of the same type as var1.


- statement;
+ statement;

where statement represents the buggy statement.

FP15. Remove Buggy Statement. Deleting entirely the buggy statement from the program. Implemented in: HDRepair, SOFix, FixMiner, CapGen, and AVATAR.

FP15.1: - ....; statement;....
FP15.2: - methodDeclaration(Arguments) {
- ....; statement;....
}
We describe in the following subsections the role and operation of TBar's performance against the state-of-the-art.

**Table 2: Change properties of fix patterns.**

<table>
<thead>
<tr>
<th>Fix Pattern</th>
<th>Change Action</th>
<th>Change Granularity</th>
<th>Bug Context</th>
<th>Change Spread</th>
</tr>
</thead>
<tbody>
<tr>
<td>FP1</td>
<td>Insert</td>
<td>statement</td>
<td>expression</td>
<td>single</td>
</tr>
<tr>
<td>FP2 (2, 3, 4, 5)</td>
<td>Insert</td>
<td>statement</td>
<td>expression, an expression returning non-primitive type data</td>
<td>single</td>
</tr>
<tr>
<td>FP3</td>
<td>Insert</td>
<td>statement</td>
<td>element access of array or collection variable</td>
<td>dual</td>
</tr>
<tr>
<td>FP4 (1, 2, 3, 4)</td>
<td>Insert</td>
<td>statement</td>
<td>any statement</td>
<td>single</td>
</tr>
<tr>
<td>FP5</td>
<td>Update</td>
<td>expression</td>
<td>class instance creation expression and clone method</td>
<td>single</td>
</tr>
<tr>
<td>FP6 (1)</td>
<td>Update</td>
<td>expression</td>
<td>conditional expression</td>
<td>single</td>
</tr>
<tr>
<td>FP6 (2)</td>
<td>Insert</td>
<td>expression</td>
<td>variable declaration expression</td>
<td>single</td>
</tr>
<tr>
<td>FP7.1</td>
<td>Update</td>
<td>expression</td>
<td>integral division expression</td>
<td>single</td>
</tr>
<tr>
<td>FP8 (1, 2, 3)</td>
<td>Update</td>
<td>expression</td>
<td>integral division expression</td>
<td>single</td>
</tr>
<tr>
<td>FP9 (1, 2)</td>
<td>Update</td>
<td>expression</td>
<td>literal expression</td>
<td>single</td>
</tr>
<tr>
<td>FP10.1</td>
<td>Update</td>
<td>expression, or statement</td>
<td>method invocation, class instance creation, constructor, or super constructor</td>
<td>single</td>
</tr>
<tr>
<td>FP10.2</td>
<td>Delete</td>
<td>expression</td>
<td>assignment or infix-expression</td>
<td>single</td>
</tr>
<tr>
<td>FP10.3</td>
<td>Insert</td>
<td>expression</td>
<td>arithmetic infix-expression</td>
<td>single</td>
</tr>
<tr>
<td>FP10.4</td>
<td>Delete</td>
<td>expression</td>
<td>instance of expression</td>
<td>single</td>
</tr>
<tr>
<td>FP11.1</td>
<td>Update</td>
<td>expression</td>
<td>return statement</td>
<td>single</td>
</tr>
<tr>
<td>FP11.2</td>
<td>Update</td>
<td>expression</td>
<td>variable expression</td>
<td>single</td>
</tr>
<tr>
<td>FP12</td>
<td>Move</td>
<td>statement</td>
<td>any statement</td>
<td>single or multiple</td>
</tr>
<tr>
<td>FP13 (1, 2)</td>
<td>Update</td>
<td>statement</td>
<td>any statement</td>
<td>single or multiple</td>
</tr>
<tr>
<td>FP14</td>
<td>Delete</td>
<td>method</td>
<td>any statement</td>
<td>multiple</td>
</tr>
</tbody>
</table>

**Figure 1: The overall workflow of TBar.**

**Table 3: Diversity of fix patterns w.r.t change properties.**

<table>
<thead>
<tr>
<th>Action Type</th>
<th># fix patterns</th>
<th>Granularity</th>
<th># fix patterns</th>
<th>Spread</th>
<th># fix patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Update</td>
<td>17</td>
<td>Expression</td>
<td>21</td>
<td>Single</td>
<td>in</td>
</tr>
<tr>
<td>Delete</td>
<td>4</td>
<td>Statement</td>
<td>17</td>
<td>Statement</td>
<td>single</td>
</tr>
<tr>
<td>Insert</td>
<td>13</td>
<td>Method</td>
<td>1</td>
<td>multiple</td>
<td></td>
</tr>
<tr>
<td>Move</td>
<td>1</td>
<td>Method</td>
<td>1</td>
<td>Statements</td>
<td>7</td>
</tr>
</tbody>
</table>

etc.) to boost automated program repair beyond the performance that a straightforward application of common fix patterns can offer. Figure 1 overviews the workflow that we have implemented in TBar. We describe in the following subsections the role and operation of each process as well as all necessary implementation details.

### 3.1 Fault Localization

Fault localization is necessary for template-based APR as it allows to identify a list of suspicious code locations (i.e., buggy statements) on which to apply the fix patterns. TBar leverages the GZoltar [9] framework to automate the execution of test cases for each buggy program. In this framework, we use the Ochiai [4] ranking metric to compute the suspiciousness scores of statements that are likely to be the faulty code locations. This ranking metric has been demonstrated in several empirical studies [56, 65, 73, 78] to be effective for localizing faults in object-oriented programs. The GZoltar framework for fault localization is also widely used in the literature of APR [18, 24, 34, 38, 39, 49, 69, 74, 76, 77], allowing for a fair assessment of TBar’s performance against the state-of-the-art.

### 3.1.2 Fix Pattern Selection

In the execution of the repair pipeline, once the fault localization process yields a list of suspicious code locations, TBar sequentially attempts to select the encoded fix patterns from its database of fix patterns for each statement in the locations list. The selection of fix patterns is conducted in a naive way based on the AST context information of each suspicious statement. Specifically, TBar sequentially traverses each node of the suspicious statement AST from its first child node to its last leaf node and tries to match each node against the context AST of the fix pattern. If a node can match any bug context presented in Table 2, a related fix pattern will be matched to generate patch candidates with the corresponding code change pattern. If the node is not a leaf node, TBar keeps traversing its children nodes. For example, if the first child node of a suspicious statement is a method invocation expression, it will be first matched with FP10. **Mutate Method Invocation Expression** fix pattern. If the children nodes of the method invocation start from a variable reference, it will be matched with FP13. **Mutate Variable** fix pattern as well. Other fix patterns follow the same manner. After all expression nodes of a suspicious statement are matched with fix patterns, TBar further matches fix patterns from statement and method levels respectively.

### 3.1.3 Patch Generation and Validation

When a matching fix pattern is found (i.e., a pattern is selected for a suspicious statement), a patch is generated by mutating the statement, then the patched program is run against the test suite. If the patched program passes all tests successfully, the patch candidate is considered as a plausible patch [58]. Once such a plausible patch is identified, TBar stops generating other patch candidates for this bug to fix bugs in a standard and practical program repair workflow [38, 39, 49, 76, 77], but does not generate all plausible patches for each bug, unlike PraPR [15]. Otherwise, the pattern selection and patch generation process is resumed until all AST nodes of buggy code are traversed. When several fix pattern contexts match one node, their actions are used for ordering: TBar prioritizes Update over Insert that is prioritized over Move. In case of multiple donor code options for a given fix pattern, the candidate patches (each generated with a specific donor code) are ordered based on the distances between donor code node and buggy code node in the AST of the buggy code file: priority is given to smaller distances. Due to space limitation, detailed steps, illustrated in an algorithmic pseudo-code, are released in the replication package.

Considering that some buggy programs have several buggy locations, if a patch candidate can make a buggy program pass a sub-set of previously failing test cases without failing any previously passing test cases, this patch is considered as a plausible sub-patch of this buggy program. TBar will further validate other patch candidates, until either a plausible patch is generated, or all patch candidates are validated, or TBar exhausts the time limitation set (i.e., three hours) for repair attempts.
If a plausible patch is generated, we further manually check the equivalence between this patch and the ground-truth patch provided by developers and available in the Defects4J benchmark. If the plausible patch is semantically equivalent to the ground-truth patch, the plausible patch is considered as correct. Otherwise, it is only considered as plausible. We offer a replication package with extensive details on pattern implementation within TBar. Source code is publicly available in the aforementioned GitHub repository.

3.2 Assessment Benchmark

For our empirical assessments, we selected the Defects4J [20] dataset as the evaluation benchmark of TBar. This benchmark includes test cases for buggy Java programs with the associated developer fixes. Defects4J is an ideal benchmark for the objective of this study, since it has been widely used by most recent state-of-the-art APR systems targeting Java program bugs. Table 4 provides summary statistics on the bugs and test cases available in the version 1.2.0 [1] of Defects4J which we use in this study.

![Table 4: Defects4J dataset information.](image)

Overall, we note that, to date, 101 Defects4J bugs have been correctly fixed by at least one APR tool published in the literature. Nevertheless, we recall that SimFix [18] currently holds the record number of bugs fixed by a single tool, which is 34.

4 ASSESSMENT

This section presents and discusses the results of repair experiments with TBar. In particular, we conduct two experiments for:

- **Experiment #1:** Assessing the effectiveness of the various fix patterns implemented in TBar. To avoid the bias that fault localization can introduce with its false positives (cf. [38]), we directly provide perfect localization information to TBar.
- **Experiment #2:** Evaluating TBar in a normal program repair scenario. We investigate in particular the tendency of fix patterns to produce more or less incorrect patches.

4.1 Repair Suitability of Fix Patterns

Our first experiment focuses on assessing the patch generation performance of fix patterns for real bugs. In particular, we investigate three research questions in Experiment #1.

<table>
<thead>
<tr>
<th>Research Questions for Experiment #1</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RQ1. How many real bugs from Defects4J can be correctly fixed by fix patterns from our taxonomy?</td>
<td></td>
</tr>
<tr>
<td>RQ2. Can each Defects4J bug be fixed by different fix patterns?</td>
<td></td>
</tr>
<tr>
<td>RQ3. What are the properties of fix patterns that are successfully used to fix bugs?</td>
<td></td>
</tr>
</tbody>
</table>

In a recent study, Liu et al. [38] reported how fault localization techniques substantially affect the repair performance of APR tools. Given that, in this experiment, the APR tool (namely TBar) is only used as a means to apply the fix patterns in order to assess their effectiveness, we must eliminate the fault localization bias. Therefore, we assume that the bug positions at statement level are known, and we directly provide it to the patch generation step of TBar, without running any fault localization tool (which is part of the normal APR workflow, see Figure 1). To ensure readability across our experiments, we denote this version of the APR system as TBar$_p$ (where $p$ stands for perfect localization). Table 5 summarizes the experimental results of TBar$_p$.

![Table 5: Number of bugs fixed by fix patterns with TBar$_p$.](image)

Among 395 bugs in the Defects4J benchmark, TBar$_p$ can generate plausible patches for 101 bugs. 74 of these bugs are fixed with correct patches. We also note that TBar$_p$ can partially fix 20 bugs with plausible patches, and 8 of them are correct. In a previous study, the kPAR [38] baseline tool (i.e., a Java implementation of the PAR [23] seminal template-based APR tool) was correctly/plausibly fixing 36/55 Defects4J bugs when assuming perfect localization.

While the results of TBar$_p$ are promising, the 79% (314/395) of bugs cannot be correctly fixed with the available fix patterns. We manually investigated these unfixed bugs and make the following observations as research directions for improving the fix rates:

- **Insufficient fix patterns.** Many bugs are not fixed by TBar$_p$, simply due to the absence of matching fix patterns. This suggests that the fix patterns collected in the literature are far from being representative for real-world bugs. The community must thus keep contributing with effective techniques for mining fix patterns from existing patches.

- **Ineffective search of fix ingredients.** Template-based APR is a kind of search-based APR [69]: some fix patterns require donor code (i.e., fix ingredients) to generate actual patches. For example, as shown in Figure 2, to apply the relevant fix pattern FF9.2, one needs to identify fix ingredient “ImageMapUtilities.htmlEscape” as the necessary in generating the patch. The current implementation of TBar limits its search space for donor code to the “local” file where the bug is localized. It is a limitation to find the correct donor code, but it reduces the risk of search space explosion. In addition, TBar leverages the context of buggy code to prune away irrelevant fix ingredients. Therefore, some bugs cannot be fixed by TBar although its fix pattern can match with code change actions. With more effective search strategies (e.g., larger search space such as fix ingredients from other projects as in [34]), there might be more chances to fix more bugs.

**RQ1:** The collected fix patterns can be used to correctly fix 74 real bugs from the Defects4J dataset. A larger portion of the dataset remains however unfixed by TBar$_p$, notably due to (1) the limitations of the fix patterns set and to (2) the naïve search strategy for finding relevant fix ingredients to build concrete patches from patterns.

Figure 3 summarizes the statistics on the number of bugs that can be fixed by one or several fix patterns. The Y-axis denotes the number of fix patterns (i.e., $n = 1, 2, 3, 4, 5,$ and $>5$) that can generate plausible patches for a number of bugs (X-axis). The legend indicates that “P” represents the number of plausible patches generated by TBar$_p$ (i.e., those that are not found to be correct). “#k”, where

---

[3] Partial fix: a patch makes the buggy program pass a part of previously failed test cases without causing any new failed test cases [10].
the importance of selecting a relevant donor code for synthesizing patches: selecting an inappropriate donor code can lead to the generation of a plausible (but incorrect) patch, which will impede the generation of correct patches in a typical repair pipeline.

Aside from fix patterns, fix ingredients collected in donor code are essential to be properly selected to avoid patches that are plausible but may yet be incorrect.

We further inspect properties of fix patterns, such as change actions, granularity, and the number of changed statements in patches. The statistics are shown in Figure 4, highlighting the number of plausible (but incorrect) and correct patches for the different property dimensions through which fix patterns can be categorized.

In summary, 86% (= 74/85) of correctly fixed bugs (74 fully and 7 partially fixed bugs) are exclusively fixed correctly by single patterns. In other words, generally, several fix patterns can generate patches that can pass all test cases but, in most cases, the bug is correctly fixed by only one pattern. This finding suggests that it is necessary to carefully select an appropriate fix pattern when attempting to fix a bug, in order to avoid plausible patches which may prevent the discovery of correct patches by halting the repair process (given that all tests are passing on the plausible patch).

We further inspect properties of fix patterns, such as change actions, granularity, and the number of changed statements in patches. The statistics are shown in Figure 4, highlighting the number of plausible (but incorrect) and correct patches for the different property dimensions through which fix patterns can be categorized.

RQ2: Some bugs can be plausibly fixed by different fix patterns. However, in most cases, only one fix pattern is adequate for generating a correct patch. This finding suggests a need for new research on fix pattern prioritization.

Table 6 details which bug is fixed by which fix pattern(s). We note that five fix patterns (i.e., FP3, FP4.3, FP5, FP7.2 and FP11.3) cannot be used to generate a plausible patch for any Defects4J bug. Two fix patterns (i.e., FP9.2 and FP12) lead to plausible patches for some bugs, but none of them is correct. It does not necessarily suggest that the aforementioned fix patterns are useless (or ineffective) in APR. Instead, two reasons can explain their performance:

- The search for donor code may be inefficient for finding relevant ingredients for applying these patterns
- The Defects4J dataset does not contain the types of bugs that can be addressed by these fix patterns.

In addition, twenty (20) fix patterns lead to the generation of correct patches for some bugs. Most of these fix patterns are involved in the generation of plausible patches (which turn out to be incorrect). Interestingly, we found the cases of six (6) fix patterns which can generate several\(^4\) patch candidates, some which being correct and others being only plausible, for the same 10 bugs (as indicated in Table 6 with ‘\(^{•}\)’). This observation further highlights

---

\(^4\)Note that, in this experiment TBar\(_n\) generates and assesses all possible patch candidates for a given pair “bug location - fix pattern” with varying ingredients.

---

4.2 Repair Performance Comparison: TBar vs State-of-the-art APR tools

Our second experiment evaluates TBar in a realistic setting for patch generation, allowing for reliable comparison against the state-of-the-art in the literature. Concretely, we investigate two research questions in Experiment #2.

<table>
<thead>
<tr>
<th>Research Questions for Experiment #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>RQ4: What performance can be achieved by TBar in a standard and practical repair scenario?</td>
</tr>
<tr>
<td>RQ5: To what extent are the different fix patterns sensitive to noise in fault localization (i.e., spotting buggy code locations)?</td>
</tr>
</tbody>
</table>

In this experiment we implement a realistic scenario, using a normal fault localization (i.e., no assumption of perfect localization as for TBar\(_n\)) on Defects4J bugs. To enable a fair comparison with performance results recorded in the literature, TBar leverages a standard configuration in the literature [38] with GZoltar [9] and Ochiai [4]. Furthermore, TBar does not utilize any additional technique to improve the accuracy of fault localization, such as crashed stack trace (used by ssFix) [4], predicate switching [50] (used by ACS) [76], or test case purification [79] (used by SimFix [18]).

With respect to the patch generation step, contrary to the experiment with TBar\(_n\) where all positions of multi-locations bugs were known (cf. Section 4.1), TBar adapts a “first-generated and first-selected” strategy to progressively apply fix patterns, one at a time, in various suspicious code locations: TBar generates a patch \(p_i\),
using a fix pattern that matches a given bug. If $p_i$ passes a subset of previously-failing test cases without failing any previously-passing test case, `TBar` selects $p_i$ as a plausible patch for the bug. Then, `TBar` continues to validate another patch $p_{i+1}$ (which can be generated by the same fix pattern on the same code entity with other ingredients, or on another code location). When $p_{i+1}$ passes a subset of test cases as $p_i$, if $p_{i+1}$ is generated for the same buggy code entity as $p_i$, $p_{i+1}$ will be abandoned; otherwise, `TBar` takes $p_{i+1}$ as another plausible patch as well. Through this process, `TBar` creates a patch set $P = \{ p_1, p_1, \ldots \}$ of plausible patches. Here, as soon as any patch can pass all the given test cases for a given bug, `TBar` takes it as a plausible patch for the given bug, which is regarded as a *fully-fixed*.

### Table 6: Defects4j bugs fixed by fix patterns.

<table>
<thead>
<tr>
<th>Bug ID</th>
<th>FP1</th>
<th>FP2</th>
<th>FP3</th>
<th>FP4</th>
<th>FP5</th>
<th>FP6</th>
<th>FP7</th>
<th>FP8</th>
<th>FP9</th>
<th>FP10</th>
<th>FP11</th>
<th>FP12</th>
<th>FP13</th>
<th>FP14</th>
<th>FP15</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-1</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>□</td>
</tr>
<tr>
<td>C-2</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>□</td>
</tr>
<tr>
<td>C-3</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>□</td>
</tr>
<tr>
<td>C-4</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>□</td>
</tr>
<tr>
<td>C-5</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>□</td>
</tr>
<tr>
<td>C-6</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>□</td>
</tr>
<tr>
<td>C-7</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>□</td>
</tr>
<tr>
<td>C-8</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>□</td>
</tr>
<tr>
<td>C-9</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>□</td>
</tr>
<tr>
<td>C-10</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>□</td>
</tr>
<tr>
<td>C-11</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>□</td>
</tr>
<tr>
<td>C-12</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>□</td>
</tr>
<tr>
<td>C-13</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>□</td>
</tr>
<tr>
<td>C-14</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>□</td>
</tr>
<tr>
<td>C-15</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>□</td>
</tr>
<tr>
<td>C-16</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>□</td>
</tr>
<tr>
<td>C-17</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>□</td>
</tr>
<tr>
<td>C-18</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>□</td>
</tr>
<tr>
<td>C-19</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>□</td>
</tr>
<tr>
<td>C-20</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>□</td>
</tr>
<tr>
<td>C-21</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>□</td>
</tr>
<tr>
<td>C-22</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>□</td>
</tr>
<tr>
<td>C-23</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>□</td>
</tr>
<tr>
<td>C-24</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>□</td>
</tr>
<tr>
<td>C-25</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>□</td>
</tr>
<tr>
<td>C-26</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>□</td>
</tr>
</tbody>
</table>

*●* indicates that the bug is correctly fixed and *□* indicates that the generated patch is plausible but not correct. *●* means that the fix pattern can generate both correct patch and plausible patch for a bug. *●* and *□* denote that the bug can be partially fixed by the corresponding fix pattern. In the last column, we provide $x/y$ numbers: $x$ is the number of fix patterns that can generate correct patches for a bug, and $y$ is the number of fix patterns that can generate plausible patches for a bug. Note that the bugs that can be plausible but incorrectly fixed by fix patterns are not shown in this table.
bug, and all $P_i \in P$ will be abandoned. Otherwise, our tool yields $P$, a set of plausible patches that can each partially fix the given bug. We run the TBar APR system against the buggy programs of the Defects4J dataset. Table 7 presents the performance of TBar in comparison with recent state-of-the-art APR tools from the literature. TBar can fix 81 bugs with plausible patches, 43 of which are correctly fixed. No other APR tool had reached this number of fixed bugs. Nevertheless, its precision (ratio of correct vs. plausible patches) is lower than some recent tools such as CapGen and SimFix which employs sophisticated techniques to select fix ingredients. Nonetheless, it is noteworthy that, despite using fix patterns cataloged in the literature, we can fix three bugs (namely Cl-86,L-47,M-11) which had never been fixed by any APR system: M-11 is fixed by a pattern found by a standalone fix pattern mining tool [35] but which was not encoded by any APR system yet. Cl-86 and L-47 are fixed by patterns that were not applied to Defects4J.

**RQ4:** TBar outperforms all recent state-of-the-art APR tools that were evaluated on the Defects4J dataset. It correctly fixes 43 bugs, while the runner-up (SimFix) is reported to correctly fix 34 bugs.

It is noteworthy that TBar performs significantly less than TBar$_p$ (43 vs. 74 correctly fixed bugs). This result is in line with a recent study [38], which demonstrated that fault localization imprecision is detrimental to APR repair performance. Table 6 summarizes information about the number of bugs each fix pattern contributed to fixing with TBar$_p$. While only 4 fix patterns did not lead to the generation of any plausible patch when assuming perfect localization. With TBar, it is the case for 13 fix patterns (see Table 8). This observation further confirms the impact of fault localization noise. We propose to examine the locations where TBar applied fix patterns to generate plausible but incorrect patches. As shown in Figure 5, TBar has made changes on incorrect positions (i.e., non-buggy locations) for 24 out of the 38 fully-fixed and 15 out of the 16 partially-fixed bugs.

**Figure 5:** The mutated code positions of plausibly but incorrectly fixed bugs.

Even when TBar applies a fix pattern to the precise buggy location, the generated patch may be incorrect. As shown in Figure 5, 14 patches that fully fix Defects4J bugs mutate the correct locations: in 3 cases, the fix patterns were inappropriate; in 2 other cases, TBar failed to locate relevant donor code; for the remaining, TBar does not support the required fix patterns.

Finally, Figure 6 illustrates the impact of fault localization performance: unfixed bugs (but correctly fixed by TBar$_p$) are generally more poorly localized than correctly fixed bugs. Similarly, we note that many plausible but incorrect patches are generated for bugs which are not well localized (i.e., several false positive buggy locations are mutated leading to plausible but incorrect patches).

**5 DISCUSSION**

Overall, our investigations reveal that a large catalogue of fix patterns can help improve APR performance. However, at the same time, there are other challenges that must be dealt with: more accurate fault localization, effective search of relevant donor code, fix pattern prioritization. While we will work on some of these research directions in future work, we discuss in this section some threats to validity of the study and practical limitations of TBar.

**5.1 Threats to Validity**

Threats to external validity include the target language of this study, i.e., Java. Fix patterns studied in this paper only cover the fix patterns targeting at Java program bugs released by the state-of-the-art pattern-based APR systems. However, we believe that most fix patterns presented in this study could be applied to other languages since fix patterns are illustrated as abstract syntax tree level. Another threat to external validity could be the fix pattern diversity. Our study may not consider all available fix patterns so far in the literature. To reduce this threat, we systematically reviewed the

<table>
<thead>
<tr>
<th>Project</th>
<th>GenProg</th>
<th>Jali</th>
<th>JMatReap</th>
<th>HDRepair</th>
<th>Nopol</th>
<th>ACS</th>
<th>ELIXIR</th>
<th>JAAD</th>
<th>Ids</th>
<th>FixGen</th>
<th>SketchFix</th>
<th>FixMiner</th>
<th>L5Repair</th>
<th>SimFix</th>
<th>KPAR</th>
<th>AVATAR</th>
<th>TBar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chart</td>
<td>0.7</td>
<td>0.6</td>
<td>1/4</td>
<td>0/2</td>
<td>1/6</td>
<td>2/2</td>
<td>4/7</td>
<td>2/4</td>
<td>3/7</td>
<td>4/4</td>
<td>6/8</td>
<td>5/9</td>
<td>3/9</td>
<td>4/8</td>
<td>3/10</td>
<td>5/12</td>
<td>9/14</td>
</tr>
<tr>
<td>Closure</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
</tr>
<tr>
<td>Long</td>
<td>0/0</td>
<td>0/0</td>
<td>0/1</td>
<td>2/6</td>
<td>3/7</td>
<td>3/4</td>
<td>8/12</td>
<td>1/8</td>
<td>5/12</td>
<td>5/5</td>
<td>3/4</td>
<td>2/3</td>
<td>8/14</td>
<td>9/13</td>
<td>1/8</td>
<td>5/11</td>
<td>5/14</td>
</tr>
<tr>
<td>Math</td>
<td>5/18</td>
<td>1/4</td>
<td>2/11</td>
<td>4/7</td>
<td>1/21</td>
<td>12/16</td>
<td>12/39</td>
<td>1/8</td>
<td>10/26</td>
<td>12/7</td>
<td>7/8</td>
<td>12/14</td>
<td>14/26</td>
<td>7/18</td>
<td>6/13</td>
<td>29/36</td>
<td>0/4</td>
</tr>
<tr>
<td>Mockito</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
</tr>
<tr>
<td>Time</td>
<td>0/12</td>
<td>0/2</td>
<td>0/0</td>
<td>0/1</td>
<td>0/3</td>
<td>1/2</td>
<td>2/3</td>
<td>0/4</td>
<td>0/4</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
</tr>
</tbody>
</table>

| FP(%)   | 18.5    | 4.3  | 17.6     | 26.1    | 14.3  | 78.5 | 63.4   | 29.9  | 33.3| 80.0   | 73.1       | 80.6     | 51.4    | 60.7   | 36.7 | 50.9  | 53.1  |

**5.1 Threats to Validity**

Threats to external validity include the target language of this study, i.e., Java. Fix patterns studied in this paper only cover the fix patterns targeting at Java program bugs released by the state-of-the-art pattern-based APR systems. However, we believe that most fix patterns presented in this study could be applied to other languages since fix patterns are illustrated as abstract syntax tree level. Another threat to external validity could be the fix pattern diversity. Our study may not consider all available fix patterns so far in the literature. To reduce this threat, we systematically reviewed the
works have shown that the adequate donor code, for some bugs, is valid for assessment. The performance achieved by TBar is naïvely matches patterns based on the AST context: it naïvely matches patterns based on the AST context selects fix patterns in a naïve way, it thus would be necessary to design a sophisticated strategy (such as bug symptom, bug type, or other information from bug reports) for fix pattern selection to reduce the noise from inappropriate fix patterns. Searching donor code for synthesis patches is another limitation of TBar, as the correct donor code for fixing some bugs is located in the code files that do not contain the bug [18, 69]. If TBar extends the donor code searching to other non-buggy code files, it will cause the search space explosion.

5.2 Limitations
TBar selects fix patterns in a naïve way, it thus would be necessary to design a sophisticated strategy (such as bug symptom, bug type, or other information from bug reports) for fix pattern selection to reduce the noise from inappropriate fix patterns. Searching donor code for synthesis patches is another limitation of TBar, as the correct donor code for fixing some bugs is located in the code files that do not contain the bug [18, 69]. If TBar extends the donor code searching to other non-buggy code files, it will cause the search space explosion.

6 RELATED WORK
Fault Localization. In general, most APR pipelines start with fault localization (FL), as shown in Figure 1. Once the buggy position is localized, ARP tools can mutate the buggy code entity to generate patches. To identify defect locations in a program, several automated FL techniques have been proposed [72]: slice-based [47, 71], spectrum-based [6, 57], statistics-based [32, 33], etc.

Spectrum-based FL is widely adopted in APR systems since they identify bug position at the statement level. It relies on the ranking metrics (e.g., Trantula [19], Ochiai [5]) to calculate the suspiciousness of each statement. GZoltar [9] and Ochiai have been widely integrated into APR systems since their effectiveness has been demonstrated in several empirical studies [56, 65, 73, 78]. As reported by Liu et al. [38] and studied in this paper, FL configuration still has a limitation on localizing bug positions. Therefore, researchers tried to enhance FL techniques with new techniques, such as predicate switching [76, 80] and test case purification [18, 79].

Patch Generation. Another key process of APR pipelines is searching for another shape of a program (i.e., a patch) in the space of all possible programs [30, 43]. If the search space is too small, it might not include the correct patches. [69]. To reduce this threat, a straightforward strategy is to expand the search space, however, which could lead to other two problems: (1) at worst, there still is no correct patch in it; and (2) the expanded search space includes more plausible patches that enlarge the possibility of generating plausible patches before correct ones [34, 69].

To improve repair performance, many APR systems have been explored to address the search space problem. Synthesis-based APR systems [42, 76, 77] explored to limit the search space on conditional bug fixes by synthesizing new conditional expressions with variables identified from the buggy code. Pattern-based APR tools [13, 17, 18, 23, 25, 29, 39–41, 62] are designed to purify the search space by following fix patterns to mutate buggy code entities with retrieved donor code. Other APR pipelines focus on specific search methods for donor code or patch synthesizing strategies, to address the search space problem, such as contract-based [10, 66], symbolic execution based [53], learning based [7, 16, 44, 59, 64, 70], and donor code searching [21, 51] APR tools. Various existing APR tools have achieved promising results on fixing real bugs, but there is still an opportunity to improve the performance; for example, mining more fix patterns, improving pattern selection and donor code retrieving strategy, exploring a new strategy for patch generation, and prioritizing bug positions.

Patch Correctness. The ultimate goal of APR systems is to automatically generate a correct patch that can resolve the program defects. In the beginning, patch correctness is evaluated by passing all test cases [23, 29, 67]. However, these patches could be overfitting [27, 58] and even worse than the bug [63]. Since then, APR systems are evaluated with the precision of generating correct patches [18, 39, 69, 76]. Recently, researchers start to explore automated frameworks that can identify patch correctness for APR systems automatically [28, 75].

7 CONCLUSION
Fix patterns have been studied in various scenarios to understand bug fixes in the wild. They are further implemented in different APR pipelines to generate patches automatically. Although template-based APR tools have achieved promising results, no extensive investigation on the effectiveness fix patterns was conducted. We fill this gap in this work by revisiting the repair performance of fix patterns via a systematic study assessing the effectiveness of a variety of fix patterns summarized from the literature. In particular, we build a straightforward template-based APR tool, TBar, which we evaluate on the Defects4j benchmark. On the one hand, assuming a perfect fault localization, TBar fixes 74/101 bugs correctly/plausibly. On the other hand, in a normal/practical APR pipeline, TBar correctly fixes 43 bugs despite the noise of fault localization false positives. This constitutes a record performance in the literature on Java program repair. We expect TBar to be established as the new baseline APR system, leading researchers to propose better techniques for substantial improvement of the state-of-the-art.

ACKNOWLEDGMENTS
This work is supported by the Fonds National de la Recherche (FNR), Luxembourg, through RECOMMEND 15/IS/10449467 and FIXPATTERN C15/IS/9964569.
REFERENCES


