# ObfusX: Routing Obfuscation with Explanatory Analysis of a Machine Learning Attack

Wei Zeng
wei.zeng@wisc.edu
University of Wisconsin–Madison
Madison, WI, USA

Azadeh Davoodi
adavoodi@wisc.edu
University of Wisconsin–Madison
Madison, WI, USA

Rasit Onur Topaloglu
rasit@us.ibm.com
IBM
Hopewell Junction, NY, USA

## ABSTRACT

This is the first work that incorporates recent advancements in "explainability" of machine learning (ML) to build a routing obfuscator called ObfusX. We adopt a recent metric—the SHAP value—which explains to what extent each layout feature can reveal each unknown connection for a recent ML-based split manufacturing attack model. The unique benefits of SHAP-based analysis include the ability to identify the best candidates for obfuscation, together with the dominant layout features which make them vulnerable. As a result, ObfusX can achieve better hit rate (97% lower) while perturbing *significantly* fewer nets when obfuscating using a via perturbation scheme, compared to prior work. When imposing the same wirelength limit using a wire lifting scheme, ObfusX performs significantly better in performance metrics (e.g., 2.4 times more reduction on average in percentage of netlist recovery).

## CCS CONCEPTS

• **Security and privacy** → **Security in hardware**; • **Hardware** → **VLSI design manufacturing considerations**; • **Computing methodologies** → **Machine learning**.

## KEYWORDS

routing obfuscation, split manufacturing, explainable artificial intelligence, machine learning

**ACM Reference Format:**
Wei Zeng, Azadeh Davoodi, and Rasit Onur Topaloglu. 2021. ObfusX: Routing Obfuscation with Explanatory Analysis of a Machine Learning Attack . In *26th Asia and South Pacific Design Automation Conference (ASPDAC '21), January 18–21, 2021, Tokyo, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3394885.3431600

## 1 INTRODUCTION

Manufacturing outsourcing of Integrated Circuits has become more common than ever before because of the high cost of fabricating high-end chips. As a result, security issues including design piracy and hardware Trojans injection may arise when an untrusted foundry is involved in manufacturing. To alleviate these problems, split manufacturing is proposed as a technique where the untrusted foundry only receives and fabricates a partial layout up to a metal layer denoted by a "split level". However, this may still not prevent an attacker to extract the full design, if the layout is not obfuscated or if the split level is too high, as suggested by [1–6].

Existing techniques on design obfuscation may be classified as two categories: placement-based and routing-based. Placement-based techniques include pin swapping [1], cell insertion [7], and cell location perturbation [2]. Routing-based techniques include routing blockage insertion [3], routing perturbation [8], and wire lifting [9]. The two techniques may also been combined, as in [10].

The key idea of design obfuscation for split manufacturing is to make an attack model fail to identify correct connections above the split level. As for the attack models for split manufacturing, Rajendran *et al.* first proposed the proximity attack [1]. Wang *et al.* proposed a more advanced network-flow-based proximity attack [2], which employs the network flow model that considers more heuristics for better attack performance. Magaña *et al.* proposed a congestion based attack [3], which redefined proximity measures based on the observation that placement and routing congestions are better indicators in large commercial designs. Most recently, Zeng *et al.* proposed a machine learning (ML) attack model [4], trained with empirically-selected layout features that reflect the hints from routing conventions.

**In this paper**, we propose a novel way to build an obfuscator for split manufacturing, based on recent advancements in the area of "explainability" of ML. We adopt a recent explanatory metric, namely the **SH**apley **A**dditive ex**P**lanation (SHAP) value [11], to analyze the ML attack model proposed in [4]. (The ML attack model is especially suitable for large commercial designs while other attack models (e.g. [2]) would take prohibitively long attack time.)

The SHAP-based analysis reveals to what extent *each* layout feature contributes to correctly predicting *each* individual unknown connection as seen by an untrusted foundry. We then exploit this information to design a SHAP-guided obfuscator against the ML attack model where only truly vulnerable connections are identified and each is obfuscated by just the necessary amount. This results in minimal perturbation to the layout as measured by increase in wirelength and number of perturbed nets. Our obfuscator (named ObfusX) is routing-based and is performed by utilizing via perturbation and wire lifting schemes. (Placement-based obfuscation was not found to be as effective by our SHAP-based analysis.)

Overall, our contributions can be summarized as follows.

- This is the first work that shows how explainability in machine learning (ML) can be used to obfuscate a design; while we focus on routing obfuscation for an ML-based split manufacturing attack, our approach is generalizable to build any obfuscator as long as an ML attack model is available.


*ASPDAC '21, January 18–21, 2021, Tokyo, Japan*

ACM ISBN 978-1-4503-7999-1/21/01…$15.00
https://doi.org/10.1145/3394885.3431600

- We demonstrate the benefits of ObfusX in identifying and focusing on the most vulnerable candidates and obfuscating each by just the right amount, thereby reducing the obfuscation overhead, while having better performance.
- Our results are compared with two prominent prior works, using not only the ML attack, but also an independent network flow-based attack from a recent work.

## 2 PRELIMINARIES

To build an obfuscator, we use an explanatory model named SHAP to break a ML-based attack. Here, we review the ML attack model used by our work and then give a brief overview of SHAP analysis.

### 2.1 ML Attack Model for Split Manufacturing

Given a metal layer as the split level, the layout is partitioned into public layers, v-pins (as termed in [4]) and private layers from low to high levels. A *split layer* refers to the topmost metal layer available to the attacker; *public layers* refer to all metal on or below the split layer and via layers in between; *private layers* are all metal layers above the split layer and the via layers in between; *v-pins* are vias connecting public and private layers. The attacker has access to the layout (cells, pins, wires, vias) in public layers and all v-pins. The goal of the split manufacturing attack is to predict the connectivity on private layers based on the available layout on public layers.

Recently, a ML-based attack model was proposed for split manufacturing in [4]. To build the ML model, for each pair of v-pins in a design, first a vector of layout "features" was extracted from the public layers. Using these features, the ML model was built based on Bagging of 10 reduced error pruning trees (REPTrees) in Weka [12]. The ML model mapped each v-pin pair with feature vector $\mathbf{x}$ to a probability $f(\mathbf{x}) \in [0, 1]$, indicating how likely the v-pin pair is a "match" (i.e. actually connected to each other on private layers).

For a pair of v-pins, the following features were extracted in [4]. (We refer the reader to [4] for more details.)

- `diffVpinX`, `diffVpinY`: The $x$- and $y$-direction differences in the locations of the two underlying v-pins, respectively.
- `manhattanVpin`: The Manhattan distance of the v-pins.
- `diffPinX/Y`, `manhattanPin`: These are similarly defined but for pins (connected to the v-pin pair) at the placement level.
- `totalWireLength`: Wirelengths of wires connected to the v-pin pair on public layers.
- `totalCellArea`, `diffCellArea`: These are sum (or diff) of the average area of output cells and that of input cells.

### 2.2 SHAP Tree Explainer for Machine Learning

In this work, we adopt a recently proposed explanatory model named SHAP [11], which explains predictions from a ML model. Let $f(\mathbf{x}_i)$ denote the ML prediction output of the $i$-th testing sample with feature vector $\mathbf{x}_i$. SHAP decomposes the model output as

$$f(\mathbf{x}_i) = \mathbb{E}[f(\mathbf{x})] + \sum_{j=1}^{M} c_{i,j}, \tag{1}$$

where $\mathbb{E}[f(\mathbf{x})]$ is the expected prediction based on all training data, and $c_{i,j}$ is the contribution of the $j$-th feature of the $i$-th testing sample, which can be positive, negative, or zero. Each $c_{i,j}$ indicates

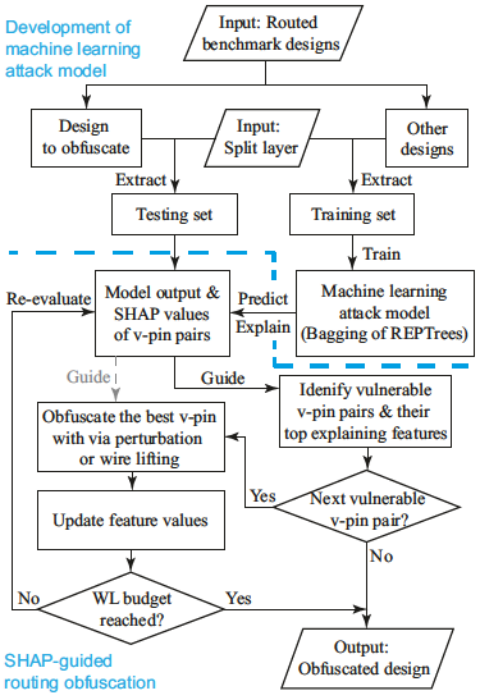


**Figure 1: Flow chart of ObfusX.**

to what extent the $j$-th feature deviates the $i$-th sample's prediction from the average. The SHAP value is proposed as an excellent candidate to compute the $c_{i,j}$s in (1). SHAP values show how each feature contributes to the model output for each testing sample.

A recent extension [13], referred to as *SHAP tree explainer*, shows that the *exact* evaluation of SHAP values can be done in polynomial time exclusively for tree-based models (which is compatible with the aforementioned ML attack model). The SHAP tree explainer does *not* assume feature independence, as feature interactions are already captured in the underlying trees. In this paper, we will use the SHAP tree explainer to analyze the vulnerability of individual v-pin pairs to the attack and use it to guide the obfuscation.

## 3 OVERVIEW OF OBFUSX

The core idea of a SHAP-guided obfuscation is to perturb the design, such that a ML attack model would perform worse. As we will show in experiments, such obfuscation also performs well under an independent, non-ML, attack model [2]. This is because both attack models are based on a similar set of design conventions in routing tools that aim to optimize the wirelength, delay, etc. A flow chart of the overall process of ObfusX is shown in Figure 1.

The upper panel shows how the ML model is developed. To generate the training set and testing set for a design to obfuscate (i.e., "target design"), we generate data samples by extracting layout features from routed designs, with the same split layer applied as will be used in manufacturing. All data samples from the target design are allocated in the testing set, which we will use to monitor the progress and performance of obfuscation. Other designs in the same benchmark suite as the target design are used to generate the

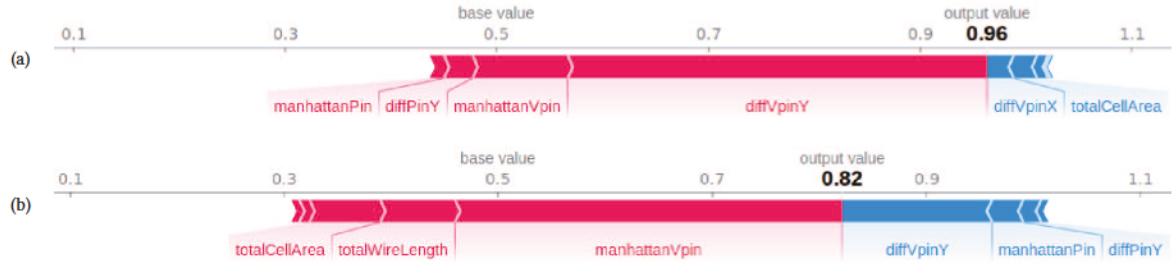


**Figure 2: SHAP force plots of two actually-connected v-pin pairs. The pink/blue bars (to the left/right of output values, respectively) quantify to what extend each layout feature positively/negatively contributes to the ML attack that predicts their connectivity. The top contributing features (longest pink bars) may vary from one v-pin pair to another. For example, `diffVpinY` is the most contributing feature in predicting (a) (longest pink bar) while it is actually the most negatively contributing feature to predicting (b) (longest blue bar).**

training set that will be used to train the attack model. As mentioned in Section 2, ObsufX uses the ML predictor in [4]. With a trained attack model, it predicts how likely each pair of (two) v-pins in the target design could be a match (i.e., are actually connected), which can be interpreted as the vulnerability of the pair to the ML attack.

To develop ObfusX, as shown in the lower panel, the ML prediction for a v-pin pair is fed to the SHAP tree explainer, which generates a set of SHAP values to explain the prediction.

Each SHAP value corresponds to an extracted feature and quantifies to what extent that feature contributes to the ML predictor for that specific v-pin pair. These SHAP values are next analyzed across all actually-connected v-pin pairs to identify the most vulnerable ones to the ML attack, along with the layout features that contribute the most to their individual vulnerabilities.

Next, the output of SHAP analysis guides the actual obfuscation which is done iteratively. ObfusX utilizes two layout perturbation techniques—via perturbation and wire lifting—each of which effectively change the routing and locations of a vulnerable v-pin pair. At each iteration, the most vulnerable v-pin pair is obfuscated if its obfuscation does not violate routing feasibility. Next, the feature vector of the obfuscated pair is updated and consequently its vulnerability is re-evaluated by the attack model (given that the layout has been slightly perturbed). ObfusX then proceeds to obfuscate the next vulnerable pair, until there is no more vulnerable pair, or a budget of wirelength (WL) overhead is reached.

## 4 SHAP ANALYSIS FOR ONE V-PIN PAIR

Before discussing the details of ObfusX, we first explain how SHAP-based analysis is performed for a single pair of connected v-pins. This helps us to illustrate the true benefits of such analysis in building ObfusX. Consider two connected v-pins from the design `superblue1` with split layer M6. The ML attack model, predicts this pair to be connected with probability 0.96 (which is a relatively high prediction indicating a successful attack if there is no obfuscation).

Figure 2(a) shows the *force plot* generated by SHAP analysis performed on the ML prediction for this pair. The color and length of pink/blue bars show the sign and magnitude of $c_{i,j}$s in Equation (1), respectively. For the pair in Figure 2(a), the analysis breaks down the prediction output of 0.96 as sum of a base value of 0.5 and a total deviation of +0.46. The pink/blue bars correspond to the features which positively/negatively contribute to the model output (i.e., with a positive/negative "force" pushing towards this 0.96 prediction). The length of the bars indicate the degree of contribution such that the sum of the lengths of pink bars (with positive sign) and blue bars (with negative sign) adds up to +0.46.

More specifically, for pair (a), among all its features, `diffVpinY` has the highest SHAP value of around +0.4 (corresponding to the length of its pink bar). Figure 2(b) shows the force plot for a second pair (b). For pair (b), we observe a different feature, i.e., `manhattanVpin` is dominant. Moreover, `diffVpinY`, which was the top feature in (a), has a negative SHAP value in (b), indicating it actually contributes negatively to the prediction of pair (b).

The above example yields the following two key observations to illustrate the unique benefits of SHAP analysis for obfuscation:

1. The vulnerable v-pin pairs can be identified as the ones which have few features with large positive SHAP values.
2. The top feature may vary across individual pairs, implying a different degree or scheme of obfuscation is needed for each.

## 5 DETAILS OF OBFUSX

The goal of SHAP-guided obfuscation is to alter the SHAP values such that there will not be any dominant feature with a high positive SHAP. It could mean that obfuscation makes originally dominant features to have a lower positive SHAP value or a negative one.

Our SHAP analysis of design `superblue1` with split layer M6, shows that for about half of the connected v-pin pairs, the SHAP value of `diffVpinY` is consistently dominant (followed by that of `manhattanVpin`). However for the other half of pairs, the distribution of SHAP values over features becomes fuzzy, which suggests that no single feature dominates the model. Such pairs (which do not have any dominant feature) do not need to be obfuscated.

For the two dominant features (`diffVpinY` and `manhattanVpin`) in the above example, Figure 3 shows the distribution of the combined contribution (i.e., sum of SHAP values) of these two top features, before and after obfuscation. This is when using ObfusX with via perturbation (which will be discussed in detail in Section 5.1). The before-obfuscation distribution is shown in blue and the

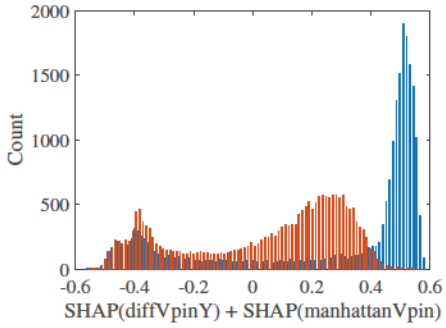


**Figure 3: Contributions of top two features `diffVpinY` and `manhattanVpin`, shown as a distribution for all connected v-pin pairs, before (blue) and after (red) obfuscation. ObfusX flattens the distribution and decreases the top contributions.**

after-obfuscation one is shown in red. As can be seen, ObfusX flattens the distribution and shifts it to the left (so it decreases the top contributions, making some less positive and some even negative).

Similar to the example of `superblue1` with split layer M6, SHAP-based analysis with the rest of the designs showed that `diffVpinY` and `manhattanVpin` are always the top two contributing features for many of the vulnerable nets when the split layer is even. (For odd split layers `diffVpinY` is replaced with `diffVpinX` because wires are preferred to route vertically on even layers and horizontally on odd layers.) The nets which did not have a dominant feature simply won't need to be obfuscated with SHAP-guided analysis. Therefore, these two features are the only ones which are utilized by ObfusX.

We note, these two dominant features are related to routing which explains our choice to obfuscate the design with routing-based techniques, i.e., via perturbation and wire lifting. However, we note, our general approach is not restricted to routing.

## 5.1 ObfusX with Via Perturbation

The procedure for via perturbation only considers perturbing v-pin pairs which are determined to be "essential". Essential v-pin pairs are a subset of all connected v-pin pairs, after disregarding trivial cases, e.g., when some v-pins connect to each other using the public layer, hence are easily identifiable by the attacker. ObfusX also ensures feasibility of the routing throughout the process without any area overhead. We first introduce the following which will be used when explaining the algorithm.

*5.1.1 Terminology.* We introduce the following terminology as shown in Figure 4(a), where the split layer is M4. Wires in all metal layers are shown as horizontal lines and vias as vertical lines.

A *driving pin* is a pin that drives other components in the net. It can be the output pin of a logic cell or that of a primary input.

A *v-pin group* consists of v-pins in the same net that connect to each other using public layers. The v-pins in the same group can be easily identified by an attacker because they are connected in public layers that are available to the attacker.

A *driving v-pin group* is a v-pin group that connects to a driving pin using public layers.

A *non-driving v-pin group* is a v-pin group that does not connect to any driving pin in public layers.

An *essential v-pin pair* $(v, v')$ consists of a pair of v-pins, where $v$ is in a non-driving v-pin group $G$, and $v'$ is in a driving v-pin group $G'$. If $G'$ has more than one v-pin, $v'$ is the closest v-pin to $v$ in $G'$.

*5.1.2 Algorithm.* We propose an algorithm that perturbs the locations of v-pins based on SHAP values of the top features `manhattanVpin` and `diffVpin`$R$ where $R$ is $X$ for odd split layers and $Y$ for even split layers. This is done iteratively, one v-pin at a time. We first calculate the SHAP values $S(i, j)$ for all essential v-pin pairs $i$ and all features $j$. Then for each essential v-pin pair $i$, we take the maximum of the SHAP values over all features $j$, i.e., $S_{\max}(i) = \max_j S(i, j)$. We only perturb v-pins that satisfy the following criteria:

- The v-pin belongs to an essential v-pin pair $p = (v, v')$, with $v$ and $v'$ in the same net. This is to avoid duplicated or invalid perturbations, e.g. perturbing the same v-pin later when a different v-pin pair is being considered.
- $S_{\max}(p) = S(p,$ `manhattanVpin`$)$ or $S_{\max}(p) = S(p,$ `diffVpin`$R)$. This ensures the essential v-pin pair $p$ is vulnerable, i.e., likely predictable with the top features.
- $S(p,$ `diffVpin`$R) \geq S(p,$ `diffVpin`$R')$, where $R' \in \{X, Y\}$ is the routing direction other than $R$. This condition ensures the effectiveness of perturbing $v$ or $v'$ in $R$ direction.
- If there are more than one non-driving v-pin group in the net of $v$ and $v'$, then $v'$ in the driving v-pin group is not eligible for perturbation and only $v$ may be perturbed. Otherwise, perturbing $v'$ may affect multiple essential v-pin pairs.

The procedures of SHAP-guided via perturbation are summarized in Algorithm 1. We maintain a list $\mathcal{L}$ of essential v-pin pairs $p = (v, v')$ sorted in decreasing order of $S_{\max}(p)$. As shown in Algorithm 1 (lines 5–6), in each iteration, we select $p$ from the top of the list, and apply trial perturbing moves (a series of "dry runs" that do not actually perturb) to each eligible v-pin in pair $p$ within a predefined small radius $r$ (detailed in Algorithm 2) to find the most efficient move $(v^*, \delta^*)$ which means to move v-pin $v^*$ by amount $\delta^*$. Efficiency of a move is defined in terms of the decrease in the model output $-\Delta f(\mathbf{x})$ and the extra WL $\Delta WL$ (as an integer). Specifically, to quantify the efficiency of a move, we define its *gain* as

$$gain = \begin{cases} -\Delta f(\mathbf{x})/\Delta WL, & \text{if } \Delta f(\mathbf{x}) < 0 \text{ and } \Delta WL \geq 1 \\ 1 - \Delta f(\mathbf{x}), & \text{if } \Delta f(\mathbf{x}) < 0 \text{ and } \Delta WL \leq 0 \\ 0, & \text{if } \Delta f(\mathbf{x}) \geq 0 \text{ or not feasible} \end{cases}, \quad (2)$$

which prioritizes moves that lead to a decrease in model output at no or low extra cost of WL. The trial perturbing is necessary as it would be difficult to estimate the routing feasibility and extra WL without any trials due to complex layout congestion. After the trial perturbing, if there is no feasible move, we remove pair $p$ from $\mathcal{L}$ (line 14), and proceed to the next v-pin pair in $\mathcal{L}$; if there is any feasible move (lines 7–11), we take the actual move that has the highest gain, update the feature vector and the SHAP values (as in Figure 1), re-check the v-pin eligibility, and go to the next iteration.

*5.1.3 Rip-up and Reroute Procedure.* To apply a perturbing move to a v-pin $v$, we rip up and reroute the wires connecting $v$ to the other components. To facilitate the rerouting procedure, we rip up $v$ and

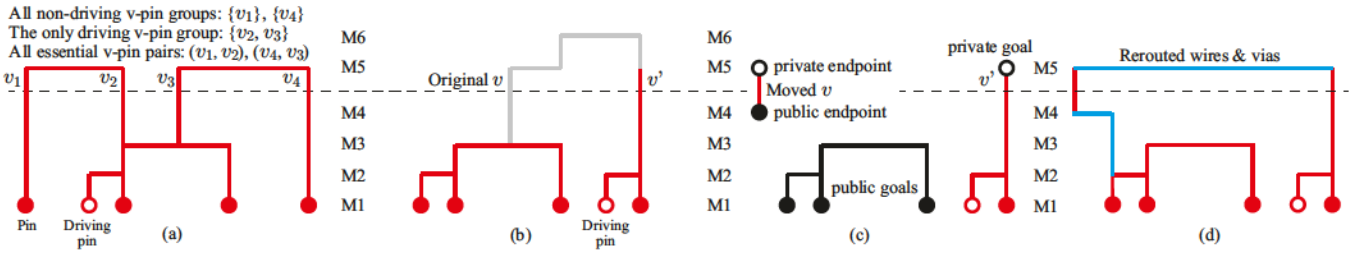


**Figure 4: (a) Illustration of terminology. (b–d) Rip up and reroute for v-pin pair $(v, v')$ when $v$ is perturbed. (b) Original wires and vias of the net containing $v$ and $v'$; the gray segments are to be removed. (c) The new location of $v$ after perturbation is identified. The unconnected parts (including both endpoints of $v$ and rerouting goals) are identified in the public layers (shown in black wires and dots) and private layers (shown in black circles). (d) The unconnected parts are reconnected (in blue).**

**Algorithm 1:** VIA-PERTURBATION $(\mathcal{L}, R, r, N)$

**Input:** $\mathcal{L}$: list of all essential v-pin pairs, $R$: perturbing direction, which is $X$ for odd split layer and $Y$ for even split layer, $r$: radius for trial perturbing, $N$: maximum number of iterations.

```
for iter ← 1 to N do
    if L is empty then
        break
    end
    for p in L in descending order of S_max(p) do
        (v*, δ*) ← TRIAL-PERTURBING(p, R, r) // Algo. 2
        if v* ≠ null then
            // take the actual move
            RIPUP-AND-REROUTE(v*, R, δ*) // Sec. 5.1.3
            Update the feature vector and SHAP values of p.
            Re-check the eligibility of both v-pins in p, and
              remove p from L if neither v-pin is eligible.
            Re-sort L by S_max.
            break  // only move one v-pin at a time
        else
            Remove p from L.
        end
    end
end
```

**Algorithm 2:** TRIAL-PERTURBING $(p, R, r)$

```
v* ← null, δ* ← null
maxGain ← 0
for eligible v in v-pin pair p do
    for δ ← −r to r do
        gain ← RIPUP-AND-REROUTE(v, R, δ) // move v in
          R-dir by δ
        if gain > maxGain then
            v* ← v, δ* ← δ
            maxGain ← gain
        end
    end
end
return (v*, δ*)  // Best v-pin to move & the amount
```

all wires connecting to $v$ that do not result in more than two connected components, while not touching any other v-pins, as shown in Figure 4(b). Then we move $v$ to the new location and identify the unconnected parts (i.e. both endpoints of $v$ and the other connected components of the net, referred to as "rerouting goals") in the public and private portions, respectively, as in Figure 4(c). Finally, we use A* search algorithm to reconnect the unconnected parts of the net in the public portion using public layers, and then reconnect for the private portion using private layers, as shown in Figure 4(d). Specifically, the routing graph $G(V, E)$ for A* search is built in three dimensions. The vertices are valid routing grids in all metal layers, and the edges are in $x$, $y$ and $z$ directions, corresponding to potential wires (in $x$ and $y$ directions) and vias (in $z$ direction) where the routing resources permit. This rip-up and reroute procedure ensures a feasible route (if possible) and optimizes the WL.

## 5.2 ObfusX with Wire Lifting

Wire lifting is the second routing-based technique in ObfusX. It moves wires from the public layers to private layers, and therefore creates more v-pins, which can make the attack more difficult.

Here, the same flow in Figure 1 is followed. However, instead of going through the v-pins connecting public and private layers as in via perturbation, we now consider the vias one layer below (i.e. the vias connecting the topmost public metal layer and the metal layer immediately below it). The goal of wire lifting is to make it most difficult for the attack model to identify the created v-pin pairs as connected, after lifting. To this end, ObfusX iteratively selects the via $v$ on this layer which, when lifted above the split layer, would create an essential v-pin pair $p$ whose maximal SHAP value $S_{\max}(p)$, is the lowest among all options of $v$.

After we select the v-pin $v$ at each iteration, we perform the wire lifting by applying the same rip up and reroute procedure to $v$ as in Section 5.1.3, except that (a) we do not move the location of $v$ after ripping up for saving WL, and (b) when rerouting with A* search, we put a higher weight on wires in public layers, so that the use of public wires is discouraged and thus extra v-pins are created.

**Table 1: Results of via perturbation with ObfusX on the ISPD'11 benchmark suite**

| Split layer | Design (#v-pins) | No obfuscation HR@0.01% / 0.1% | [4] HR@0.01% / 0.1% | [4] ΔWL% | [4] PN% / PV% | [4] $t_{CPU}$ (h) | ObfusX HR@0.01% / 0.1% | ObfusX ΔWL% | ObfusX PN% / PV% | ObfusX $t_{CPU}$ (h) |
|---|---|---|---|---|---|---|---|---|---|---|
| M6 | sb1 (44486) | 23.79 / 63.33 | 2.19 / 11.58 | 3.03 | 99.83 / 99.58 | 3.86 | 0.52 / 6.12 | 0.55 | 66.57 / 36.01 | 3.28 |
| | sb5 (60034) | 29.47 / 63.96 | 5.75 / 20.38 | 4.09 | 96.81 / 91.75 | 7.13 | 4.34 / 15.46 | 0.67 | 55.62 / 30.08 | 5.30 |
| | sb10 (89846) | 31.84 / 64.34 | 10.24 / 28.31 | 4.52 | 92.45 / 79.77 | 7.75 | 9.37 / 23.93 | 0.71 | 46.49 / 23.96 | 8.05 |
| | sb12 (80816) | 33.01 / 75.58 | 8.23 / 24.78 | 3.31 | 97.70 / 90.12 | 6.46 | 4.32 / 11.67 | 0.64 | 73.87 / 37.12 | 5.45 |
| | sb18 (36026) | 20.06 / 66.11 | 4.27 / 16.55 | 2.64 | 98.91 / 94.35 | 2.88 | 2.16 / 8.68 | 0.67 | 63.02 / 34.27 | 2.06 |
| | Average | 27.63 / 66.66 | 6.14 / 20.32 | 3.52 | 97.14 / 91.11 | 5.62 | 4.14 / 13.17 | 0.65 | 61.11 / 32.29 | 4.83 |
| M4 | sb1 (150510) | 49.82 / 68.33 | 6.46 / 25.37 | 9.50 | 99.79 / 93.91 | 9.00 | 1.70 / 24.08 | 2.14 | 65.23 / 35.26 | 18.90 |
| | sb5 (179844) | 38.78 / 60.40 | 7.54 / 23.84 | 9.86 | 96.94 / 87.87 | 11.48 | 3.03 / 23.35 | 1.87 | 51.43 / 28.09 | 18.41 |
| | sb10 (200896) | 33.50 / 60.21 | 13.16 / 37.36 | 8.53 | 91.38 / 73.21 | 15.05 | 9.81 / 36.54 | 1.31 | 38.81 / 19.55 | 17.19 |
| | sb12 (173294) | 47.07 / 71.52 | 9.01 / 22.40 | 7.61 | 98.61 / 92.32 | 13.48 | 4.42 / 17.39 | 1.12 | 65.32 / 32.81 | 18.09 |
| | sb18 (86658) | 29.83 / 59.89 | 5.15 / 17.89 | 6.43 | 99.37 / 95.29 | 4.26 | 1.87 / 10.95 | 1.53 | 57.00 / 30.80 | 7.18 |
| | Average | 39.80 / 64.07 | 8.26 / 25.37 | 8.39 | 97.22 / 88.52 | 10.65 | 4.17 / 22.46 | 1.59 | 55.56 / 29.30 | 15.95 |

## 6 EXPERIMENTAL RESULTS

We obtained the source code of the ML attack from [4], used the shap library for Python for SHAP analysis, and implemented all procedures of ObfusX in C++. Experiments were done on a Linux workstation with an Intel 16-core 3.60 GHz CPU and 64 GB memory.

### 6.1 Via Perturbation with ObfusX

We first show in Table 1 the performance of via perturbation with ObfusX using five designs in ISPD'11 benchmark suite that are also used in [3, 4, 9]. We obtain routed overflow-free designs from [4], to which we apply the proposed SHAP-based via perturbation, with parameter $r = 3\times$ routing grid size. We compare the performance and the cost of obfuscation with the via perturbation technique proposed in [4]. This is based on the same ML attack model[1].

We use the following metrics to evaluate the performance and the cost of an obfuscation[2].

- Hit rate (HR) at X%: For a v-pin $v$, we first identify the top X% of other v-pins $u$ which have the highest ML model output for essential v-pin pair $(v, u)$. These v-pins are predicted by ML to most likely be the match for $v$. We call it a "hit" of $v$ if its real matching v-pin is among the v-pins identified above. We then report the average percentage of hits of all v-pins $v$ in the design. We report this metric with X = 0.01 and 0.1. (As a point of reference, X = 0.1 results in up to 89 v-pins identified on split layer M6, or up to 200 v-pins on split layer M4 in these designs. The total number of v-pins is quite large as reported in the first column of the table.) *A lower HR means better defense.*
- WL overhead (ΔWL%): percentage of increase in WL after the obfuscation. *Lower is better.*
- Perturbed nets (PN%): number of perturbed nets divided by total number of nets that *contain any v-pin. Lower is better.*
- Perturbed v-pins (PV%): number of perturbed v-pins divided by the number of v-pins in the design. *Lower is better.*
- Total runtime of obfuscation using one CPU core ($t_{CPU}$).

[1]Note that the popular network flow attack model [2] takes prohibitively long time to run on these designs and hence is not applicable here.

[2]Note that the functions of standard cells are not available in ISPD'11 benchmark. Therefore metrics related to circuit outputs (e.g. Hamming distance (HD), output error rate (OER)) are not applicable.

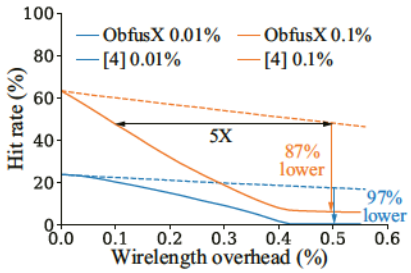


**Figure 5: Comparison of tradeoff in HR vs WL in superblue1.**

Several observations can be made from the results in Table 1. **First**, the HR of the ML model for 0.01% and 0.1% v-pin lists drops drastically after obfuscation; for ObfusX it drops from 28% and 67% to 4% and 13%, respectively, better than the HR reductions with in [4]. **Second**, the WL overhead of ObfusX is less than 1/5 of that with [4]. **Third**, with ObfusX, only around 30% of v-pins and 60% of nets (that contain v-pins) are finally perturbed, compared to nearly-all nets and v-pins when perturbed with [4].

To observe the tradeoff between performance and cost of obfuscation, we plot in Figure 5 the curves of HR and WL overhead with ObfusX and [4], respectively. Compared to [4], ObfusX achieves 87% and 97% lower HR in 0.1% and 0.01% v-pin lists, respectively, for the same WL overhead of 0.5%, or is 3–5× more efficient in WL overhead for the same reduction of HR.

### 6.2 Wire Lifting with ObfusX

We show in Table 2 the performance and cost of wire lifting with ObfusX ($r = 5\,\mu m$) on ISCAS'85 benchmark designs, which are often used in related work, and compare them with [8]. The layouts are obtained from the authors of [8].

For this benchmark, we use the network flow attack model [2] which is obtained from the authors. Note that this is *not* a ML-based attack model and is *not* used to build ObfusX. Since the split layer for each design is not explicitly reported in [8], we tried to identify it by matching the number of nets on private layers with the number reported in [8]. ObfusX was applied on six designs for which we were able to identify the split layer, with WL budget equal to the reported WL overhead in [8]. The obfuscated layouts are

**Table 2: Results of wire lifting with ObfusX with the ISCAS'85 benchmark suite**

| Design | #Nets | No obfuscation | | | [8] | | | | ObfusX | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | PNR% | OER% | HD% | PNR% | OER% | HD% | ΔWL% | PNR% | OER% | HD% | ΔWL% | $t_{CPU}$ (min) |
| c880 | 252 | 100.0 | 0.0 | 0.0 | 91.7 | 99.9 | 18.0 | 4.3 | **85.3** | **100.0** | **23.3** | **3.4** | 2.4 |
| c2670 | 607 | 95.8 | 99.9 | 7.0 | 87.1 | 100.0 | 14.0 | 4.4 | **77.8** | 100.0 | **23.5** | **3.2** | 7.0 |
| c3540 | 638 | 97.2 | 95.4 | 18.2 | 93.5 | 100.0 | 33.4 | 2.5 | **84.5** | 100.0 | **38.2** | 2.5 | 18.4 |
| c5315 | 997 | 98.7 | 98.7 | 4.3 | 95.0 | 100.0 | 18.1 | 1.7 | **88.9** | 100.0 | **23.2** | 1.7 | 13.6 |
| c6288 | 1921 | 99.8 | 36.8 | 3.0 | 98.6 | 100.0 | 42.1 | 1.8 | **95.3** | 100.0 | **45.3** | 1.8 | 14.1 |
| c7552 | 1041 | 99.6 | 69.5 | 1.6 | 95.3 | 100.0 | 20.3 | 2.2 | **87.5** | 100.0 | **27.2** | 2.2 | 12.7 |
| Avg. | | 98.5 | 66.7 | 5.7 | 93.5 | 100.0 | 24.3 | 2.8 | **86.5** | **100.0** | **30.1** | 2.5 | 11.4 |
| Comparing to "No obfus." | | | | | -5.0 | +33.3 | +18.6 | | **-12.0** | **+33.3** | **+24.4** | | |

converted to Verilog and their functional equivalency with original designs is verified with Synopsys Formality. For these designs, we use the following metrics to evaluate the performance and cost of an obfuscation.

- Percentage of netlist recovery (PNR) given in [9]: percentage of correctly reconstructed nets. This quantifies how well the attack can recover the whole design. *Lower is better.*
- Output error rate (OER): probability that there is any error bit in outputs of the reconstructed circuit. *Higher is better.*
- Hamming distance (HD) between outputs of the original and the reconstructed circuits. *Closer to 50% is better.*
- WL overhead (ΔWL%): percentage of increase in WL after the obfuscation. *Lower is better.*
- Total runtime of obfuscation using one CPU core ($t_{CPU}$).

We derive OER and HD from 100,000 runs of Monte Carlo simulations with ModelSim. OER and HD of the original design and [8], and the WL overhead of [8] are quoted from [8]. PNR of the original design and [8] are derived by definition, based on the design layouts and the reported numbers in [8].

As can be seen in Table 2, with reasonable computing time of 11 minutes on average, ObfusX reaches 100% for OER, and achieves better obfuscation in the reduction of PNR (12% vs 5% on average, or 2.4× better) and the increase in HD (24.4% vs 18.6% on average, or 31% better), with the same or less WL overhead compared to [8]. Note that the reported results of [8] come from a (best) combination of three obfuscation techniques including wire lifting and via perturbation for matching and non-matching v-pins, whereas in our results wire lifting is applied alone. In fact, our wire lifting and via perturbation techniques are orthogonal to each other. Therefore, they may be combined for potentially better performance.

We were not able to make a fair comparison with another related work [9] because the original layouts of [9] are likely to be very different from ours and were not made available. (The layouts in [9] are generated using all 10 metal layers, whereas our layouts from [8] only occupy 5–9 lower metal layers.)

*In summary*, for obfuscation with via perturbation, ObfusX is able to achieve a lower hit rate (indicating better obfuscation) while perturbing *significantly fewer* nets and vias in the design, with significantly lower wirelength. When the same wirelength limit is imposed during wire lifting, ObfusX performs significantly better in performance metrics (PNR and HD with equally good OER).

## 7 CONCLUSIONS

We presented ObfusX, a routing obfuscator for split manufacturing which incorporated SHAP-based analysis to explain a machine learning attack. The unique benefits of ObfusX were in its ability to identify the best candidate nets for obfuscation together with the layout features which make them most vulnerable when subjected to an attack. As a result, it achieved better performance than prior work while perturbing significantly fewer nets and with significantly lower wirelength during via perturbation. It also achieved significantly better performance than prior work if the same wirelength limit was imposed during wire lifting.

## ACKNOWLEDGMENTS

This research was supported by Award No. 1812600 from National Science Foundation and by Task No. 2845 from Semiconductor Research Corporation.

## REFERENCES

[1] J. Rajendran, O. Sinanoglu, and R. Karri, "Is split manufacturing secure?" in *DATE*, 2013, pp. 1259–1264.
[2] Y. Wang, P. Chen, J. Hu, G. Li, and J. Rajendran, "The cat and mouse in split manufacturing," *IEEE Trans. VLSI Syst.*, vol. 26, no. 5, pp. 805–817, 2018.
[3] J. Magaña, D. Shi, and A. Davoodi, "Are proximity attacks a threat to the security of split manufacturing of integrated circuits?" *IEEE Trans. VLSI Syst.*, vol. 25, no. 12, pp. 3406–3419, 2017.
[4] W. Zeng, B. Zhang, and A. Davoodi, "Analysis of security of split manufacturing using machine learning," *IEEE Trans. VLSI Syst.*, vol. 27, no. 12, pp. 2767–2780, 2019.
[5] Y. Wang, T. Cao, J. Hu, and J. Rajendran, "Front-end-of-line attacks in split manufacturing," in *ICCAD*, 2017, pp. 1–8.
[6] W. Xu, L. Feng, J. Rajendran, and J. Hu, "Layout recognition attacks on split manufacturing," in *ASPDAC*, 2019, pp. 45–50.
[7] K. Xiao, D. Forte, and M. M. Tehranipoor, "Efficient and secure split manufacturing via obfuscated built-in self-authentication," in *HOST*, 2015, pp. 14–19.
[8] Y. Wang, P. Chen, J. Hu, and J. Rajendran, "Routing perturbation for enhanced security in split manufacturing," in *ASPDAC*, 2017, pp. 605–610.
[9] S. Patnaik, J. Knechtel, M. Ashraf, and O. Sinanoglu, "Concerted wire lifting: Enabling secure and cost-effective split manufacturing," in *ASPDAC*, 2018, pp. 251–258.
[10] S. Patnaik, M. Ashraf, J. Knechtel, and O. Sinanoglu, "Raise your game for split manufacturing: Restoring the true functionality through BEOL," in *DAC*, 2018, pp. 140:1–140:6.
[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *NeurIPS*, 2017, pp. 4765–4774.
[12] E. Frank, M. A. Hall, and I. H. Witten, *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, 4th ed. Morgan Kaufmann, 2016.
[13] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," *arXiv:1802.03888*, 2018.