Supplementary Information

Divide and conquer： Machine learning accelerated design of lead-free solder alloys with high strength and high ductility

Qinghua Wei [&,1], Bin Cao[&,1,2], Hao Yuan[&,1], Youyang Chen[1], Kangdong You[1], Shuting Yu[1], Tixin Yang[1], Ziqiang Dong*[,1], Tong-Yi Zhang*[,1,2,3]

[1]Materials Genome Institute, Shanghai University, Shanghai, 200444, China

[2]Advanced Materials Thrust, Hong Kong University of Science and Technology (Guangzhou), Guangzhou, 511400, Guangdong, China

[3]Guangzhou Municipal Key Laboratory of Materials Informatics, Guangzhou, 511400, Guangdong, China

& These authors contributed equally to this work and should be considered co-first authors.

Corresponding authors

Correspondence to: Tong-Yi Zhang, Ziqiang Dong

zhangty@shu.edu.cn (Tong-Yi Zhang); zqdong@shu.edu.cn (Ziqiang Dong);

Supplementary Methods

Introduction of TCGPR

Tree-Classifier for Gaussian process regression model (TCGPR, github.com/Bin-Cao/TCGPR) is a data preprocessing algorithm that leverages Gaussian correlation among data. The goal of TCGPR is to maximize the Gaussian correlation by screening outliers and partitioning a messy dataset into few correlated clusters. To achieve this goal, the TCGPR uses Gaussian radial kernel function (RBF) embedded in Gaussian process regression model (GPR) integrated with leave-one-out cross validation (LOOCV) to fit and evaluate an initial dataset.

A Gaussian radial kernel function is defined as

$$K(X_{i'}, X_j) = \exp\left(-\frac{\|X_i - X_j\|^2}{2(\vartheta)^2}\right), \tag{1}$$

where $\|\cdot\|$ represents the Euclidean distance of argument vector, $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ are datum argument vectors. The length scale ϑ also implies the correlation among tested data.

Consider a dataset containing n data (X_i, y_i) $(i = 1,2,\cdots,n)$ with single objective. The GPR model is trained with leave-one-out cross validation (LOOCV), which yields in n GPR models with n optimized length scales ϑ_i and n predictions of $\hat{y}_i$. With the n optimized length scales ϑ_i, we calculate their standard deviation ϑ_{std} and mean $\bar{\vartheta}$, and with the n predictions $\hat{y}_i$ and the corresponding n real values of y_i, we calculate the Pearson correlation coefficient R between $\hat{y}_i$ and y_i. Then, the global Gaussian messy factor (GGMF) is proposed to evaluate the messiness for a given dataset, which is defined as,

$$GGMF = \omega(1 - R) + \frac{\vartheta_{std}}{\bar{\vartheta}}, \tag{2}$$

where ω is a weight constant, default as $\omega = 2$. Supplementary Eq. (2) indicates that the smaller the GGMF is, the more correlation the dataset will be. A threshold of GGMF $(GGMF)_c$ is given in TCGPR to control the level of correlation.

The process of data infilling starts from an initial seed D_0^{seed} of size n, and usually n ≥ 3. For an original dataset D of size N, there are C_N^n potential seeds $D_d^{seed}(d = 1,\cdots,C_N^n)$. Using Supplementary Eq.(2), we can obtain the initial seed D^{seed} from

$$D_0^{seed} = \underset{D_d^{seed}}{\operatorname{argmin}}(GGMF). \tag{3}$$

The rest data of the original dataset D after extracting the data in D_0^{seed} form a dataset, termed as D_0^{rest}. Next, TCGPR adds data from D_0^{rest}, one-by-one, into the seed D_0^{seed}. At any step of t, we have

$$D_{t+1}^{seed} = D_t^{seed} \cup (X_j, y_j), \left((X_j, y_j) \in D_{N-t}^{rest}\right), \quad (4)$$

where $t = 0,1, \cdots N - n$, and the TCGPR uses Supplementary Eq. (2) to calculate the $GGMF$ value on the D_t^{seed}. If the the $GGMF$ value is smaller than $(GGMF)_c$, the datum is accepted as a new member of the growing cluster, otherwise as an outlier. Iterately conducting Supplementary Eq. (4) and Supplementary Eq. (2) separates the original messy dataset into a coherently correlated cluster and another massy subset. Then, repeating the seed selection and data infilling process selects more coherently correlated clusters and outliers. More information on settings and functions of TCGPR are provided at the Open-Source platform : github.com/Bin-Cao/TCGPR.

Supplementary Discussion

Comparison results of TCGPR with other clustering algorithms

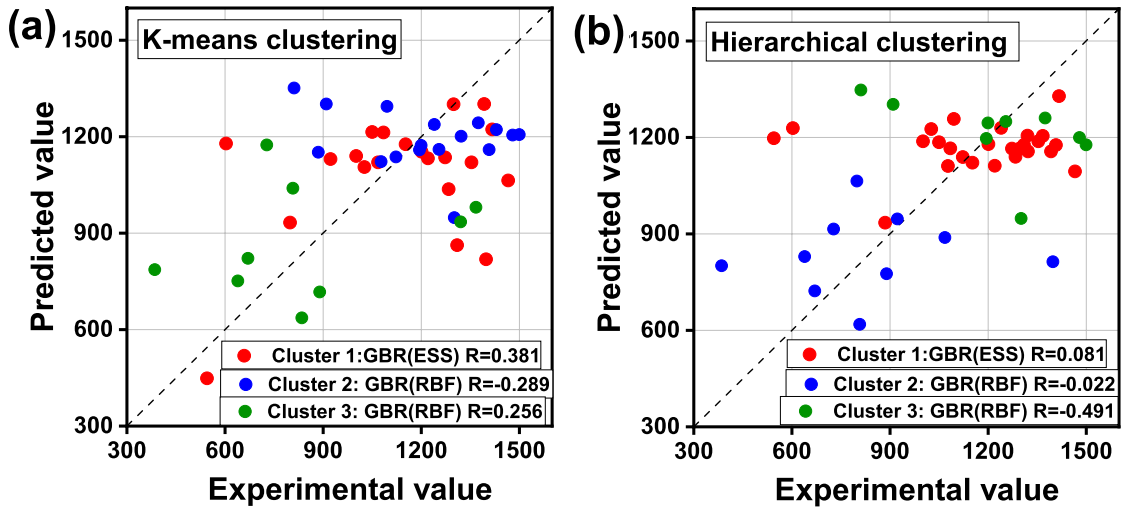
To demonstrate the advantages of the proposed TCGPR model, we compared it with K-means and Hierarchical clustering algorithms using the experimental dataset of 47 SAC387-based alloys. K-means clustering aims to partition samples into k clusters such that each sample belongs to the nearest cluster mean or centroid. With a pre-set number of clusters, the centroid positions are iteratively updated until convergence. The Hierarchical clustering is a method, with a pre-set number of clusters, for recursively clustering similar data points together in a hierarchical manner, resulting in a tree-like structure that represents the relationships between individual data points at different levels of granularity.

As a baseline, we firstly establish a GPR model with the original 47 data and obtain LOOCV prediction results of joint objective (Y), which results in R = 0.33, as described in the main text and shown in Figure 2(c). The dividing and prediction results with TCGPR are described in main text as well and the results are summarized in Supplementary Table 1.

In dividing the original dataset using either the K-means or Hierarchical clustering

algorithm, we set the number of clusters as 3 for the purpose of comparison. The clusters 1, 2, and 3 of the K-means results consist of 21, 17, and 9 data, respectively, while the clusters 1, 2, and 3 of Hierarchical clustering consist of 26, 11, and 10 data, respectively. TCGPR uses both information of features and objective and the data in each divided cluster follows the same statistical distribution, which is learned by a GPR model. The K-means and Hierarchical clustering use only the feature information and therefore the divided cluster is hard to be learned by a ML model, in comparing with the TCGPR.

GPR models are developed with the kernel functions of Gaussian radial basis function (RBF) kernel, Exp-Sine-Squared (ESS) kernel, and Dot-Product kernel and the best GPR model with the maximal LOOCV-R is selected for each of the 6 clusters. Supplementary Figures 1(a) and 1(b) show the LOOCV predictions of joint objective of the three GPR models with optimal kernels with using of K-means clustering and Hierarchical clustering, respectively. Table S1 summarizes all of the dividing and prediction results described above. Obviously, the divide and conquer TCGPR model significantly improves the LOOCV R after dividing the original dataset into subsets, whereas the clustering algorithms of K-means and hierarchical fail to achieve the same level of improvements, thereby demonstrating the advantage of the proposed TCGPR.

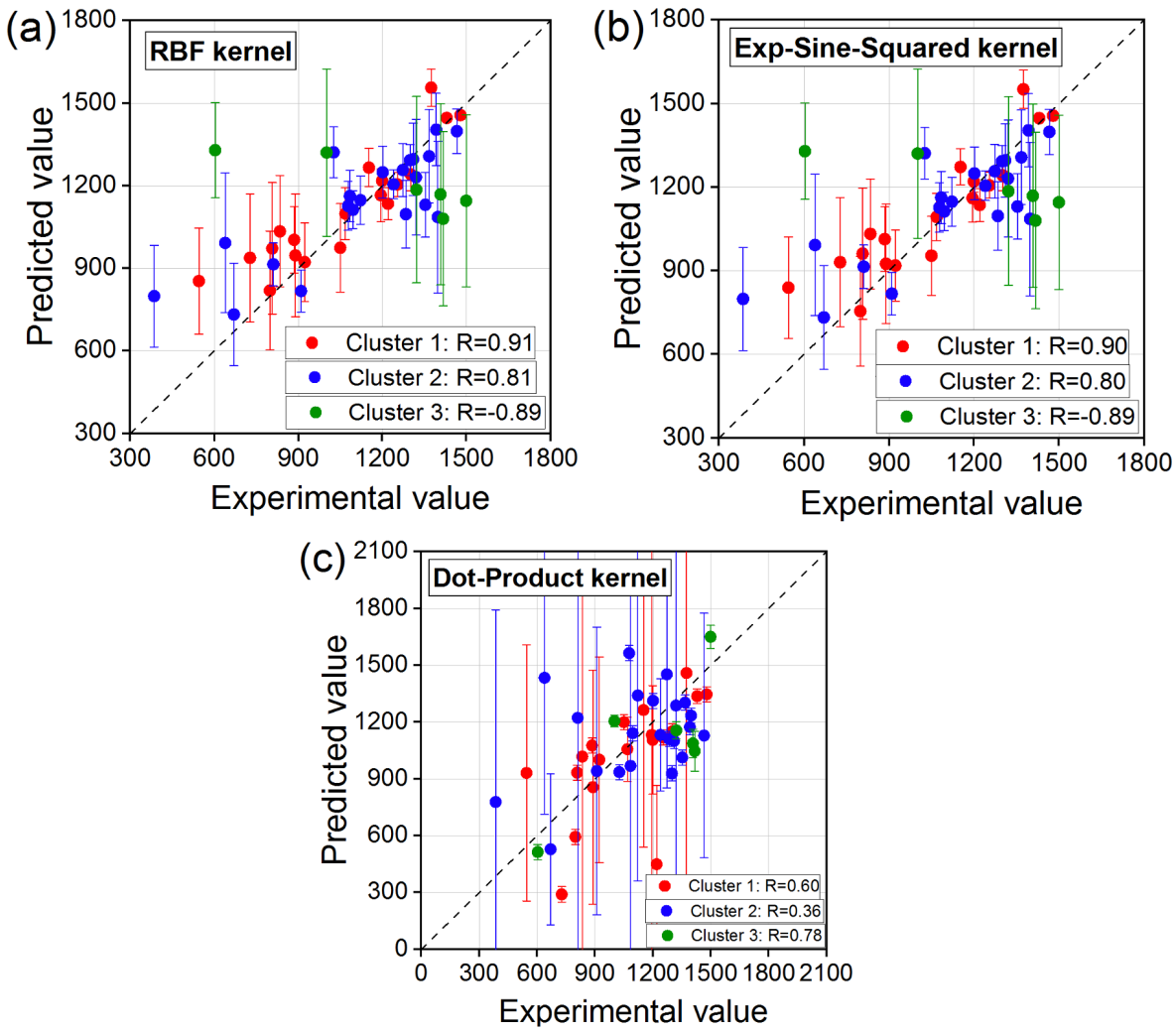


Supplementary Figure 1. The predictive value of the joint objective (MPa) versus the experimental value of three GPR models, with using (a) K-means clustering and Hierarchical clustering.

Supplementary Table 1. Summary of the results of TCGPR, K-means and Hierarchical clustering algorithms on the dataset of 47 SAC387-based alloys.

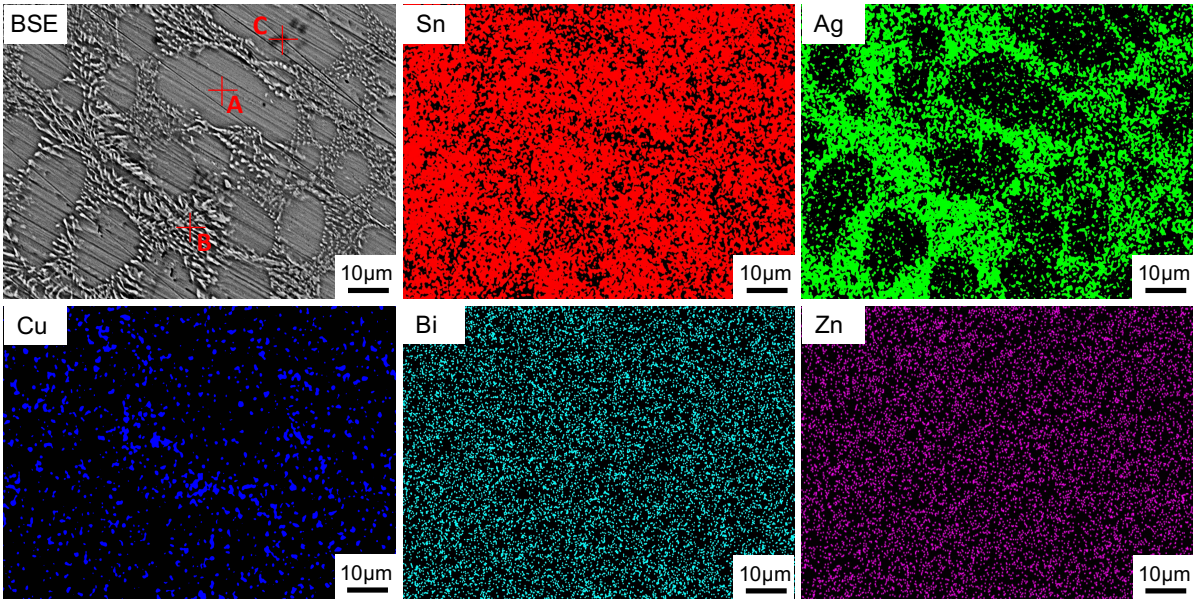
Algorithm	Cluster	Data size	Prediction model	LOOCV-R
TCGPR	Cluster 1	19	GPR(BRF)	0.91
	Cluster 2	22	GPR(BRF)	0.81
	Cluster 3	6	GPR(Dot-Product)	0.78
K-means	Cluster 1	21	GPR(Exp-Sine-Squared)	0.381
	Cluster 2	17	GPR(RBF)	-0.289
	Cluster 3	9	GPR(BRF)	0.256
Hierarchical clustering	Cluster 1	26	GPR(Exp-Sine-Squared)	0.081
	Cluster 2	11	GPR(BRF)	-0.022
	Cluster 3	10	GPR(BRF)	-0.491

GPR prediction with different kernel functions



Supplementary Figure 2. The LOOCV prediction of GPR models using different kernel functions versus the experimental value of the joint objective (Y) in three subsets: (a) Gaussian radial basis function (RBF) kernel, (b) Exp-Sine-Squared kernel and (c) Dot-Product kernel. The error bar denoted the prediction variance.

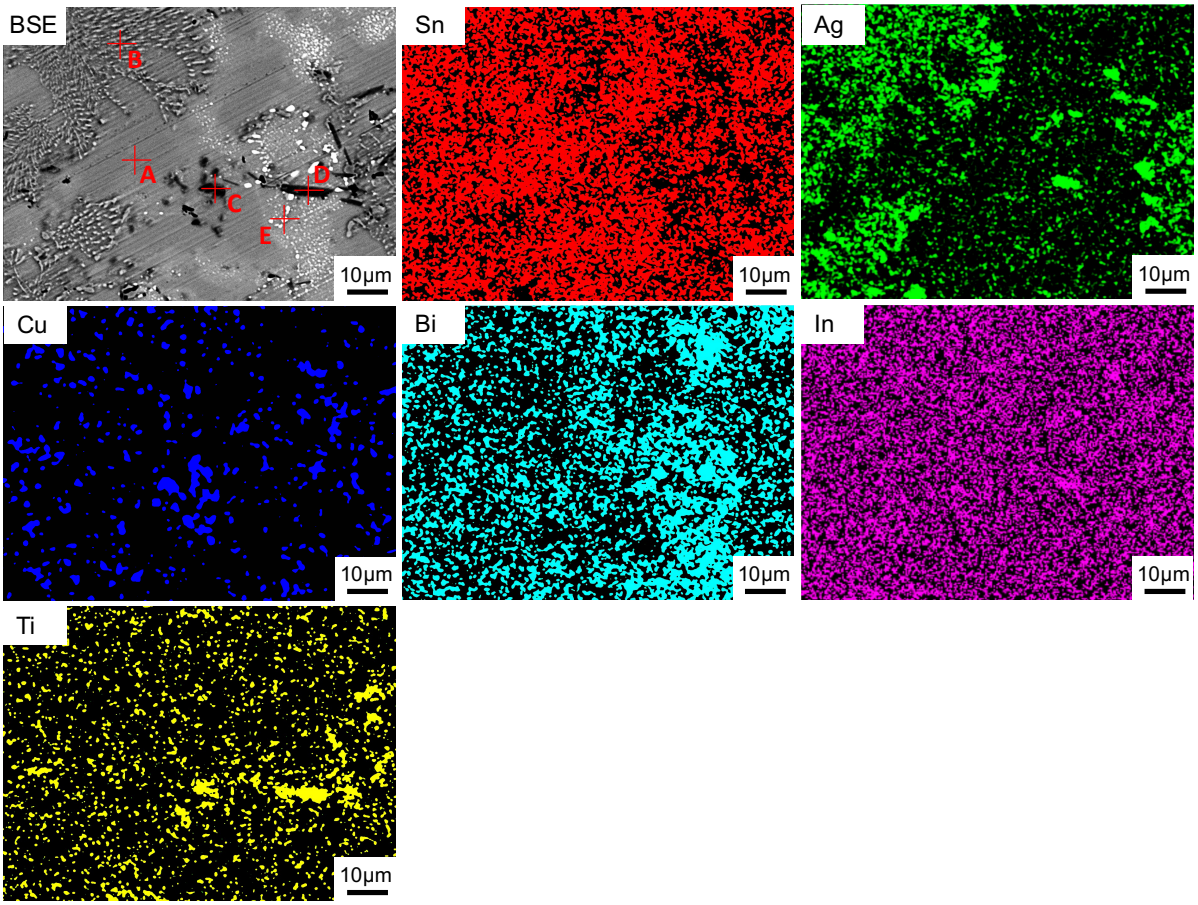
Energy Dispersive Spectroscopy (EDS) results of alloys



Supplementary Figure 3. Energy Dispersive Spectroscopy (EDS) elemental mapping results of $Sn_{94.8}Ag_{3.8}Cu_{0.7}Bi_{0.5}Zn_{0.2}$ (S1-POI) solder alloy.

Supplementary Table 2. EDS quantitative analysis results of $Sn_{94.8}Ag_{3.8}Cu_{0.7}Bi_{0.5}Zn_{0.2}$ (S1-POI).

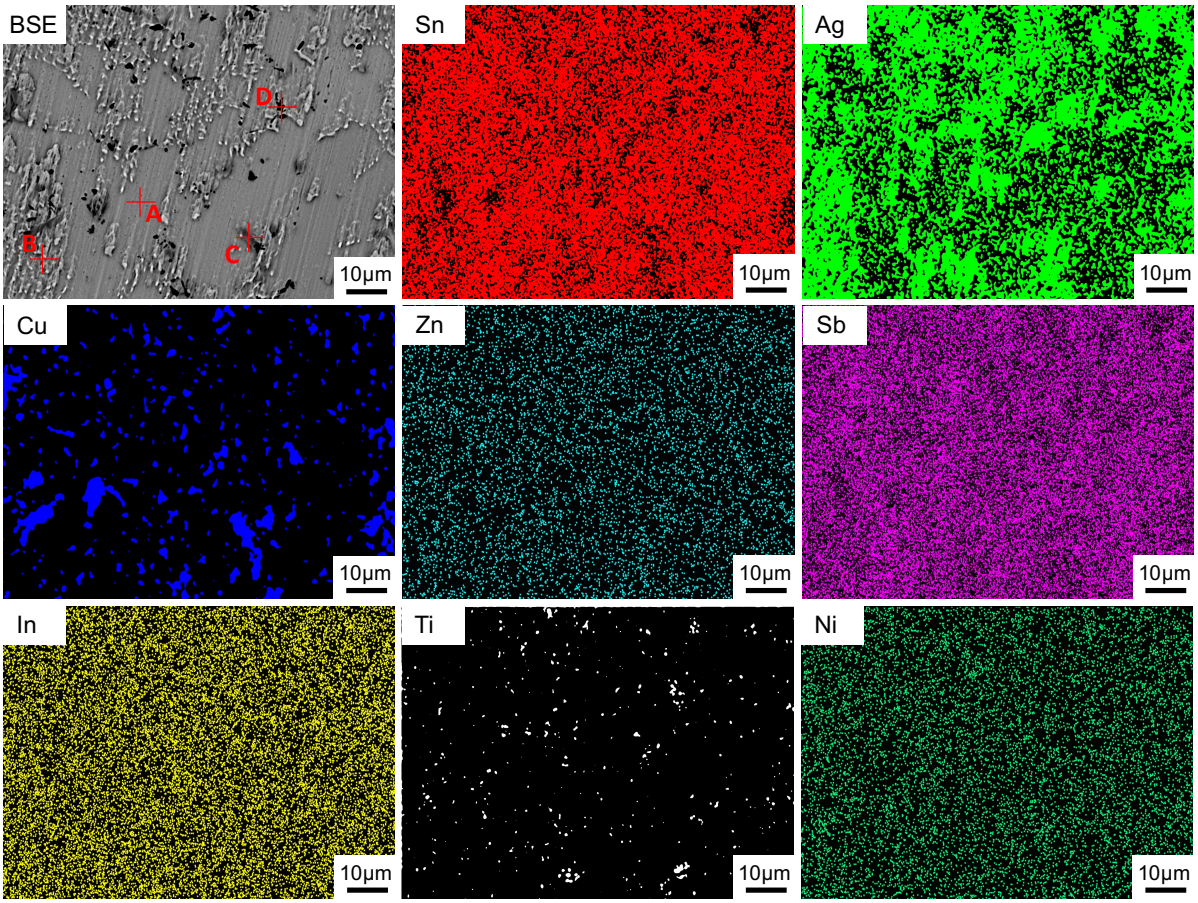
Point	Sn(at.%)	Ag(at.%)	Cu(at.%)	Bi(at.%)	Zn(at.%)	Phase
A	98.18	0	0	0.67	1.15	Sn-based solid solution
B	54.90	45.10	0	0	0	Ag_3Sn
C	48.55	0	51.45	0	0	Cu_6Sn_5



Supplementary Figure 4. EDS elemental mapping results of $Sn_{90.8}Ag_{3.8}Cu_{0.7}Bi_3In_{1.5}Ti_{0.2}$ (S2-POI) solder alloy.

Supplementary Table 3. EDS quantitative analysis results of $Sn_{90.8}Ag_{3.8}Cu_{0.7}Bi_3In_{1.5}Ti_{0.2}$ (S2-POI).

Point	Sn(at.%)	Ag(at.%)	Cu(at.%)	Bi(at.%)	In(at.%)	Ti(at.%)	Phase
A	96.89	0	0	0.62	2.49	0	Sn-based solid solution
B	42.98	51.37	0	0	5.66	0	$Ag_3(Sn, In)$
C	46.58	0	52.27	0	1.15	0	$Cu_6(Sn, In)_5$
D	64.30	0	0	0	1.73	33.97	$Ti_2(Sn, In)_3$
E	0	0	0	100	0	0	Precipitates of Bi



Supplementary Figure 5. EDS elemental mapping results of $Sn_{92.6}Ag_{3.8}Cu_{0.7}Zn_{0.6}Sb_{1.5}In_{0.5}Ti_{0.2}Ni_{0.1}$ (S3-POI) solder alloy.

Supplementary Table 4. EDS quantitative analysis results of $Sn_{92.6}Ag_{3.8}Cu_{0.7}Zn_{0.6}Sb_{1.5}In_{0.5}Ti_{0.2}Ni_{0.1}$ (S3-POI).

Point	Sn(at.%)	Ag(at.%)	Cu(at.%)	Zn(at.%)	Sb(at.%)	In(at.%)	Ti(at.%)	Ni(at.%)	Phase
A	90.79	0	0	1.40	2.59	2.24	0	2.98	Sn-based solid solution
B	55.97	40.09	0	0	0.68	3.26	0	0	$Ag_3(Sn, In, Sb)$
C	40.79	0	41.76	0	1.33	2.28	0	13.83	$(Cu, Ni)_6(Sn, In, Sb)_5$
D	59.51	0	0	0	1.23	4.74	34.52	0	$Ti_2(Sn, In, Sb)_3$

Supplementary Notes

Supplementary Table 5. 47 data of SAC387 based alloys.

No.	Sn (wt.%)	Ag (wt.%)	Cu (wt.%)	Bi (wt.%)	Zn (wt.%)	Sb (wt.%)	In (wt.%)	Ti (wt.%)	Ni (wt.%)	Al (wt.%)	Elongation (%)	UTS (MPa)	Y	Cluster
1	94.5	3.8	0.7	1	0	0	0	0	0	0	25.9	57.1	1478.9	1
2	94.3	3.8	0.7	1	0	0	0	0	0.1	0.1	29.0	49.3	1429.7	1
3	95.5	3.8	0.7	0	0	0	0	0	0	0	30.0	45.8	1374.0	1
4	93.7	3.8	0.7	1	0	0	0	0	0	0.8	24.0	54.2	1300.8	1
5	94.1	3.8	0.7	0	0	0	0	0.6	0.8	0	23.4	53.6	1253.9	1
6	91.6	3.8	0.7	3	0.4	0.5	0	0	0	0	17.0	71.8	1219.8	1
7	93.6	3.8	0.7	0.5	0	0	0	0.6	0.8	0	20.5	58.5	1199.0	1
8	93.1	3.8	0.7	1	0	1	0	0.4	0	0	14.6	81.8	1194.3	1
9	92.3	3.8	0.7	3	0	0	0	0	0.2	0	12.4	92.9	1152.0	1
10	90.4	3.8	0.7	3.5	0	1.5	0	0.1	0	0	11.8	90.5	1067.9	1
11	91.6	3.8	0.7	2	0	0	1.5	0	0.4	0	14.6	71.9	1049.7	1
12	89.5	3.8	0.7	3	0	3	0	0	0	0	9.7	95.1	922.5	1
13	87.5	3.8	0.7	5	0	3	0	0	0	0	9.4	94.6	889.2	1
14	92.7	3.8	0.7	0	0	1	1	0.6	0.2	0	14.6	60.6	884.8	1
15	89.3	3.8	0.7	5	0	0	0	0.4	0	0.8	9.7	86.0	834.2	1
16	86.8	3.8	0.7	3	0	0	5	0	0.7	0	10.9	74.0	806.9	1
17	88.5	3.8	0.7	0	0	5	1.5	0.4	0.1	0	11.0	72.6	798.6	1
18	86.8	3.8	0.7	1.5	0	2.5	4	0.7	0	0	9.2	79.1	727.6	1
19	90.2	3.8	0.7	0	0.2	3	1.5	0.6	0	0	8.0	68.1	544.8	1
20	91.3	3.8	0.7	3	0	0	1	0.2	0	0	17.1	85.7	1465.5	2
21	88.5	3.8	0.7	0	0	5	2	0	0	0	17.9	78.1	1398.0	2
22	91	3.8	0.7	1.5	0	2	0	0.6	0.4	0	17.2	80.9	1391.9	2
23	89.8	3.8	0.7	2.5	0	0	3.1	0	0.1	0	17.0	80.4	1366.8	2
24	90	3.8	0.7	1.5	0	2.5	0	0.7	0.8	0	16.7	81.0	1353.0	2
25	88.5	3.8	0.7	3.5	0	0.5	3	0	0	0	14.9	88.6	1320.1	2
26	91.3	3.8	0.7	0	1	1.5	1.5	0.2	0	0	20.4	64.2	1309.7	2
27	91.7	3.8	0.7	3	0	0	0.5	0	0	0.3	16.5	78.7	1298.6	2
28	91.5	3.8	0.7	1	0	3	0	0	0	0	19.6	65.5	1283.8	2
29	91	3.8	0.7	1.5	0	0	2	0.6	0.4	0	17.3	73.6	1272.7	2
30	92.3	3.8	0.7	0	0.6	1.5	1	0	0.1	0	21.3	58.2	1239.7	2
31	91	3.8	0.7	1.5	0	0	2.5	0.1	0.4	0	24.1	49.8	1201.0	2
32	92.7	3.8	0.7	0	0	1.5	1	0.2	0.1	0	20.0	56.1	1122.0	2
33	92.7	3.8	0.7	0	0.6	1	1	0	0.2	0	20.9	52.4	1095.2	2
34	92.2	3.8	0.7	3	0	0	0	0	0.2	0.1	13.5	80.3	1084.1	2
35	92.5	3.8	0.7	0	0.4	1	1	0.6	0	0	17.1	63.0	1077.3	2
36	90.7	3.8	0.7	1	0	3	0	0	0.8	0	15.2	67.5	1026.0	2
37	94.1	3.8	0.7	1	0	0	0	0	0	0.4	16.1	56.5	909.7	2
38	94.3	3.8	0.7	1	0	0	0	0	0.2	0	14.4	56.3	810.7	2

39	89.7	3.8	0.7	5	0	0	0	0	0	0.8	8.2	81.7	669.9	2
40	87.1	3.8	0.7	5	0	3	0	0	0.4	0	6.3	101.4	638.8	2
41	88.7	3.8	0.7	5	0	1	0	0	0	0.8	4.9	78.6	385.1	2
42	94.1	3.8	0.7	1	0	0	0	0	0.4	0	25.8	58.1	1499.0	3
43	91	3.8	0.7	1.5	0	2.5	0	0.1	0.4	0	18.4	77.0	1416.8	3
44	92.3	3.8	0.7	0	0.6	1.5	1	0.1	0	0	23.3	60.4	1407.3	3
45	92.9	3.8	0.7	0	0.4	1	1	0	0.2	0	24.2	54.6	1321.3	3
46	90.9	3.8	0.7	1	0.4	2	1	0.1	0.1	0	14.7	68.2	1000.8	3
47	91.5	3.8	0.7	2.5	0	1	0.5	0	0	0	7.5	80.7	603.2	3