# Data-driven catalyst optimization for stereodivergent asymmetric synthesis of α-allyl carboxylic acids by iridium/boron hybrid catalysis

Hongyu Chen[1,†], Shigeru Yamaguchi[2,†,*], Yuya Morita[1], Hiroyasu Nakao[1], Xiangning Zhai[1], Yohei Shimizu[1,3], Harunobu Mitsunuma[1,*] and Motomu Kanai[1,*]

[1] Graduate School of Pharmaceutical Sciences, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan.

[2] RIKEN Center for Sustainable Resource Science, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan.

[3] Present address: Department of Chemistry, Faculty of Science, Hokkaido University, Kita 10 Nishi 8, Kita-ku, Sapporo, Hokkaido 060-0810, Japan.

E-mail; shigeru.yamaguchi.hw@riken.jp; h-mitsunuma@mol.f.u-tokyo.ac.jp; kanai@mol.f.u-tokyo.ac.jp

†These authors contributed equally: H. Chen, S. Yamaguchi.

**Asymmetric catalysis enabling divergent control of multiple stereocenters remains challenging in synthetic organic chemistry. While machine learning-based optimization of molecular catalysis is an emerging approach, data-driven catalyst design to achieve stereodivergent asymmetric synthesis producing multiple reaction outcomes, such as constitutional selectivity, diastereoselectivity, and enantioselectivity, is unprecedented. Here, we report the straightforward identification of asymmetric two-component iridium/boron hybrid catalyst systems for α-*C*-allylation of carboxylic acids. Structural optimization of the chiral ligands for iridium catalysts was driven by molecular field-based regression analysis with a dataset containing overall 32 molecular structures. The catalyst systems enabled selective access to all the possible isomers of chiral carboxylic acids bearing contiguous stereocenters. This stereodivergent asymmetric catalysis is applicable to late-stage structural modifications of drugs and their derivatives.**

The design and development of functional molecules rely heavily on a researcher's intuition and time- and labor-intensive experimental trials and errors. A data-driven approach is an emerging tool to facilitate these processes. Typical data-driven molecular design approaches use a large number of dataset molecules for machine learning and/or virtual libraries to conduct *in silico* screening[1]. Exploration of chemical space on the basis of mathematical models constructed by machine learning techniques facilitates the identification of molecules exhibiting the desired properties. These approaches are applied to predict catalytic reactions using several hundred to more than 10 million pieces of compound/reaction data[2-8]. Another useful data science approach for exploring chemical space in molecular catalysis is linear regression analysis based on free energy relationships[9] developed by Sigman and coworkers[10-13]. By analyzing a relatively small amount of data (typically less than 100 samples), this method extracts important information about how molecular properties (so-called descriptors), such as electronic and steric properties, influence the reaction outcomes. The information obtained by the regression analysis is useful to search for desired molecules in chemical space. Fujita and Winkler highlighted such two types of supervised learning using molecular descriptors to predict molecular properties; i.e., models generated for predictive purposes relying on machine learning methods using large, chemically diverse datasets, and for mechanistic interpretation using small sets of chemically similar molecules[14]. Regression-based data-driven design of molecular catalysis, particularly asymmetric catalysis, is a rapidly growing research field[10-11,15-17]. Denmark's group recently reported a machine learning-based workflow for predicting chiral catalysts showing remarkably higher enantioselectivity than those in the training set[4]. Sigman's group demonstrated the construction of a regression model for predicting enantioselectivity in a range of transformations that proceeds through similar mechanisms[13]. The above two recent milestones in this area focused on predicting the enantioselectivity of products with one stereocenter. Because the stereochemistry of organic compounds can drastically influence important properties such as the biological activity and physical characteristics, however, precise catalyst control over the selectivity of products bearing multiple stereocenters is a critical issue in modern synthetic chemistry[18]. Therefore, the application of data science to the design and optimization of chiral catalysts to control multiple reaction outcomes, such as constitutional selectivity, diastereoselectivity, and enantioselectivity, for stereodivergent synthesis of all desired stereoisomers is highly important, but also extremely challenging. To achieve stereodivergent synthesis, at least four reaction outcomes (enantioselectivity and diastereoselectivity in both diastereomers) must be controlled by optimizing the catalyst structures.

## Results and Discussion

**Target reaction.** Asymmetric dual catalysis comprising two independent asymmetric catalysts to promote carbon–carbon bond-forming reactions is one of the most useful strategies for stereodivergent asymmetric synthesis[19,20]. Carreira reported pioneering work on stereodivergent dual catalysis in the $\alpha$-*C*-allylation of aldehydes by combining iridium (Ir) and amine catalyst (Fig. 1a)[19]. This approach allowed access to all the possible stereoisomers from the same starting materials under identical reaction conditions, except for the chirality of each asymmetric catalyst (Fig. 1a). Following this pioneering work, stereodivergent Ir-catalyzed allylic substitutions with prochiral enolates derived from ketones[21-23], esters[24-29] and amides[30] were reported. As a target reaction of our data-driven catalyst optimization, we chose the catalytic asymmetric migratory α-*C*-allylation of allyl esters **1** to afford α-allyl carboxylic acids **2**, using a combination of a chiral Ir complex catalyst[31-33] and a chiral boron (B) complex catalyst[34] (Fig. 1b). Previously, we developed asymmetric Pd/B hybrid catalysis for synthesis of linear α-*C*-allyl carboxylic acids containing an α-quaternary stereocenter (Fig. 1b, dotted square). We anticipated that use of an Ir catalyst instead of the Pd catalyst would afford branched carboxylic acid **2** containing contiguous α-quaternary–β-tertiary stereocenters[19,20,35,36] (Fig. 1b), which are difficult to synthesize themselves, but versatile for synthesizing various important molecules. The boron complex catalyst generates chiral boron enolate species through the chemoselective activation of carboxylate, which is formed by Ir-catalyzed ionization of **1**. The chiral boron enolate attacks the chiral Ir-π-allyl complex to stereodivergently afford products **2**. Despite recent significant progress, previous stereodivergent α-*C*-allylation of carbonyl compounds was limited to reactions affording β-aryl-substituted products ($R^3$ = Ar in Fig. 1a )[20-30]. Indeed, when Carreira's conditions were applied to an aliphatic allylic alcohol ($R^3$ = Pr), desired β-aliphatic-substituted product was not obtained at all (see Supplementary Information section 9). Thus, the development of an asymmetric Ir/B dual catalytic system that can produce β-aliphatic-substituted carboxylic acids is suitable for testing our strategy.

As the first attempt, we used phosphoramidite ligand **L1** for the Ir catalyst[31,32]. Reactions from **1Et** using boron ligands ***S*** and ***R*** provided (2*R*,3*R*)-**2Et** and (2*S*,3*R*)-**2Et** as the major isomers, respectively (Fig. 1c). Although the reactions exhibited excellent enantioselectivity (≥97% ee), the constitutional selectivity (branch (b)/linear (l) [b/l]) and diastereomer ratio (dr) were not satisfactory in reactions using ligands ***S*** (1.7/1 b/l, 1.9/1 dr) and ***R*** (2.0/1 b/l, 5.2/1 dr). Thus, we performed a data-driven optimization of the Ir-catalyst structures to improve the four reaction outcomes (b/l and dr in reactions using ligand ***S*** or ***R***), while retaining high enantioselectivity.

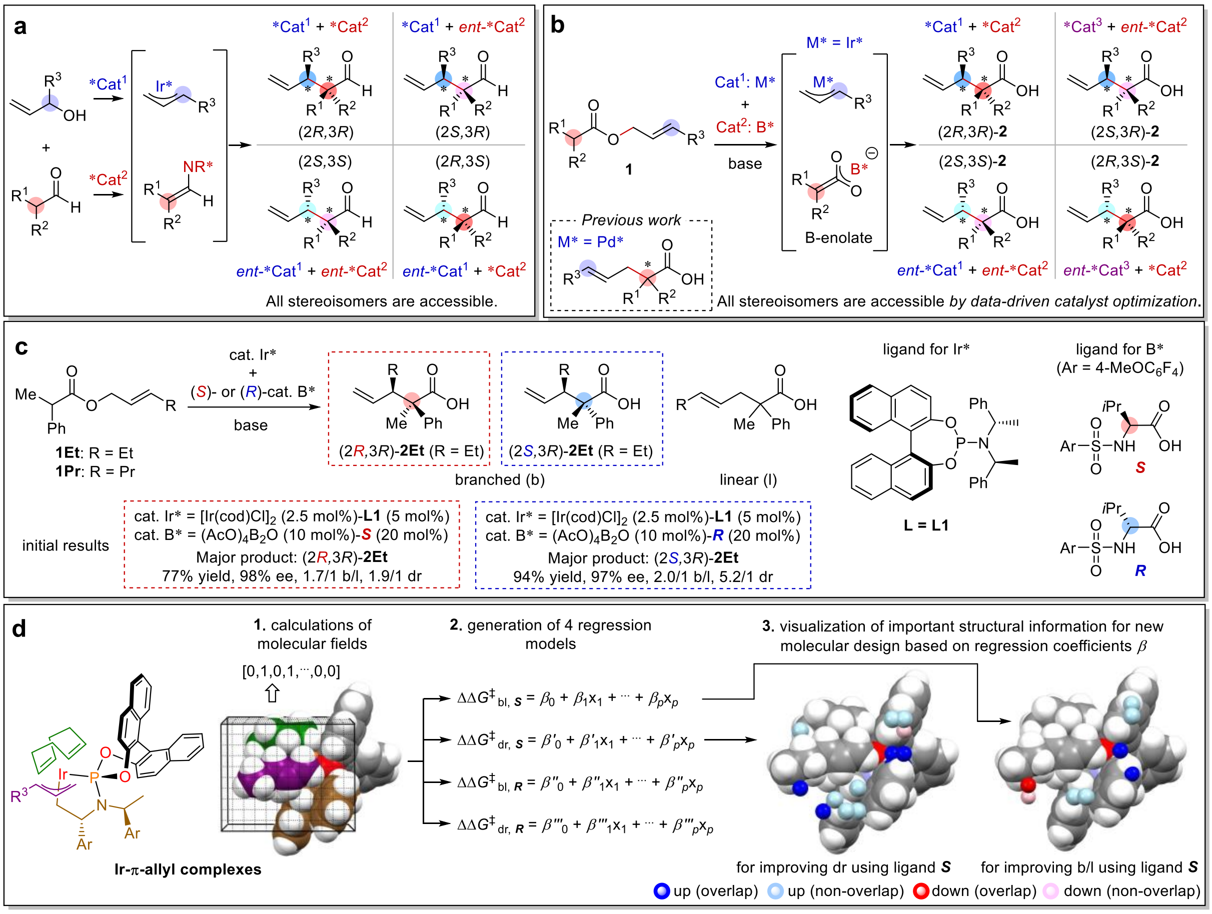

a
*Cat1 + *Cat2
*Cat1 + ent-*Cat2
*Cat1
Ir*
*Cat2
NR*
(2R,3R)
(2S,3R)
(2S,3S)
(2R,3S)
ent-*Cat1 + ent-*Cat2
ent-*Cat1 + *Cat2
All stereoisomers are accessible.
b
M* = Ir*
Cat1: M*
Cat2: B*
base
1
B-enolate
*Cat1 + *Cat2
*Cat3 + ent-*Cat2
(2R,3R)-2
(2S,3R)-2
(2S,3S)-2
(2R,3S)-2
ent-*Cat1 + ent-*Cat2
ent-*Cat3 + *Cat2
Previous work
M* = Pd*
All stereoisomers are accessible by data-driven catalyst optimization.
c
1Et: R = Et
1Pr: R = Pr
cat. Ir*
+
(S)- or (R)-cat. B*
base
(2R,3R)-2Et (R = Et)
(2S,3R)-2Et (R = Et)
branched (b)
linear (l)
ligand for Ir*
L = L1
ligand for B*
(Ar = 4-MeOC6F4)
S
R
initial results
cat. Ir* = [Ir(cod)Cl]2 (2.5 mol%)-L1 (5 mol%)
cat. B* = (AcO)4B2O (10 mol%)-S (20 mol%)
Major product: (2R,3R)-2Et
77% yield, 98% ee, 1.7/1 b/l, 1.9/1 dr
cat. Ir* = [Ir(cod)Cl]2 (2.5 mol%)-L1 (5 mol%)
cat. B* = (AcO)4B2O (10 mol%)-R (20 mol%)
Major product: (2S,3R)-2Et
94% yield, 97% ee, 2.0/1 b/l, 5.2/1 dr
d
Ir-π-allyl complexes
1. calculations of molecular fields
[0,1,0,1,···,0,0]
2. generation of 4 regression models
ΔΔG‡bl, S = β0 + β1x1 + ··· + βpxp
ΔΔG‡dr, S = β'0 + β'1x1 + ··· + β'pxp
ΔΔG‡bl, R = β''0 + β''1x1 + ··· + β''pxp
ΔΔG‡dr, R = β'''0 + β'''1x1 + ··· + β'''pxp
3. visualization of important structural information for new molecular design based on regression coefficients β
for improving dr using ligand S
for improving b/l using ligand S
up (overlap)
up (non-overlap)
down (overlap)
down (non-overlap)

**Fig. 1 | Molecular field analysis of Ir/B hybrid catalysis for stereodivergent asymmetric migratory α-*C*-allylation**. **a**, Carreira's stereodivergent asymmetric Ir/organo dual catalysis[19]. Inverting the chirality of a catalyst allowed for the access to each stereoisomer of product aldehydes. **b**, Asymmetric Ir/B dual hybrid catalysis developed in this study, providing branched carboxylic acids stereodivergently. Previous Pd/B dual catalysis (*17*) afforded linear products shown in dotted square. **c**, The initial results for the catalytic asymmetric α-*C*-allylation using a combination of a chiral iridium complex catalyst (cat. Ir*) and a chiral enolate-forming boron catalyst (cat. B*). Reaction conditions: cat. Ir* ([Ir(cod)Cl]$_2$ (2.5 mol%)–**L1** (5 mol%)) and cat. B* ((AcO)$_4$B$_2$O (10 mol%)–***S*** or ***R*** (20 mol%)) were mixed with **1Et** in the presence of DBU (1.5 equiv) and LiCl (1.0 equiv) in THF for 12 h at 50 °C. **d**, Outline of MFA-based data-driven optimization of the Ir catalysts. **1**. Calculations of 0,1 vectors (steric indicator fields) from the structures of Ir-π-allyl complexes. Indicator fields are calculated at each unit cell of a grid space. The unit cells that include the van der Waals radii of any atoms are counted as 1, or are otherwise counted as 0 (see Supplementary fig. S1). An Ir-π-allyl complex comprising **L1** (R$^3$ = Et) is shown in a CPK model: the cod ligand, the allyl (propenyl) group, and the bis(1-phenylethyl)amine moiety are shown in green, purple, and brown, respectively. **2**. Correlating the molecular fields and ΔΔ$G^‡$ values (kcal/mol) using a machine learning technique to generate 4 regression models. Logarithms of constitutional and diastereomeric ratios (b/l and dr) (ΔΔ$G^‡$ = –*RT*log(b/l or dr)) were employed as target variables, which correspond to energy differences between the transition states that lead to each isomer (Curtin-Hammett principle[56]). *p* denotes the number of descriptors. **3**. Visualization of important structural information for constitutional selectivity and diastereoselectivity on the Ir-π-allyl complexes from the regression coefficients, followed by the design of chiral phosphoramidite ligands based on the visualized guidelines. Dark blue points indicate that substituents there increased selectivity (dr or b/l) [up (overlap)]. Light blue points indicate that introducing substituents there will increase selectivity [up (non-overlap)]. Dark red points indicate that substituents there decreased selectivity [down (overlap)]. Light red points indicate that introducing substituents there will decrease selectivity [down (non-overlap)].

**Data-driven molecular design method.** For data-driven optimization of the catalyst structures in this complicated asymmetric catalysis, we employed a 3D-QSSR (quantitative structure-selectivity relationships) approach[15], which was originally developed in the field of medicinal chemistry[37] and introduced to the field of asymmetric catalysis by Lipkowitz's[38] and Kozlowski's[39] groups. Among the terms that represent 3D-QSSR, we used molecular field analysis (MFA). MFA is a regression analysis between the reaction outcomes and molecular fields calculated from three-dimensional molecular structures. From the coefficients of the constructed regression models, important structural information for the reaction outcomes can be visualized on the molecular structures. In our previous study on MFA in a relatively simple asymmetric catalysis affording one stereocenter (fluorination of a keto ester)[40], intermediates in the enantio-determining step consisting of catalysts and substrates were employed for calculations of descriptors (molecular fields). Our MFA method enabled extraction and visualization of the structural information, leading to the design of substrate molecules with improved enantioselectivity. Here, we applied the MFA method to catalyst optimization in more complicated stereodivergent asymmetric synthesis. Historically, applications of a data analysis method for a specific molecular property to evaluations and optimizations of other properties have expanded the research field. For example, Hansch and Fujita et al. utilized the extended Hammett equation to predict biological activities[41,42], which eventually led to the development of the QSAR (Quantitative Structure-Activity Relationships) field[43,44]. Our research direction is along such trends.

Because the Ir-π-allyl complex (Fig. 1b and 1d) is the hub intermediate in the selectivity-determining step of the present Ir/B dual catalysis[45] (Ir complexes and allyl groups correspond to catalysts and substrates, respectively), we utilized this complex for MFA. Importantly, since the Ir-π-allyl complex is the common intermediate in determining all the four reaction outcomes (b/l and dr in reactions using ligand ***S*** or ***R***)[46-48], one set of molecular fields calculated from the complex can be used to analyze four sets of the target variables as shown in Fig. 1d. This characteristic enables facile comparison of constitutional selectivity- and diastereoselectivity-determining factors visualized on an identical intermediate structure (Fig. 1d). Although our data analysis does not employ B-enolates for descriptor calculations, the experimental selectivity data involve the information how chiral B-enolates and Ir-π-allyl complexes interact each other when the reactions proceed. Therefore, the structural information visualized on Ir-π-allyl complexes (see below) include the information of B-enolates. It is worth noting that we can use this MFA without knowing structural details of chiral B-enolates. The obtained guidelines will lead to the efficient design of new catalysts that will simultaneously improve multiple reaction outcomes. Furthermore, the information affords insights into selectivity induction mechanisms without performing transition state density functional theory calculations (see Supplementary figs. S19 and S20)[46-48].

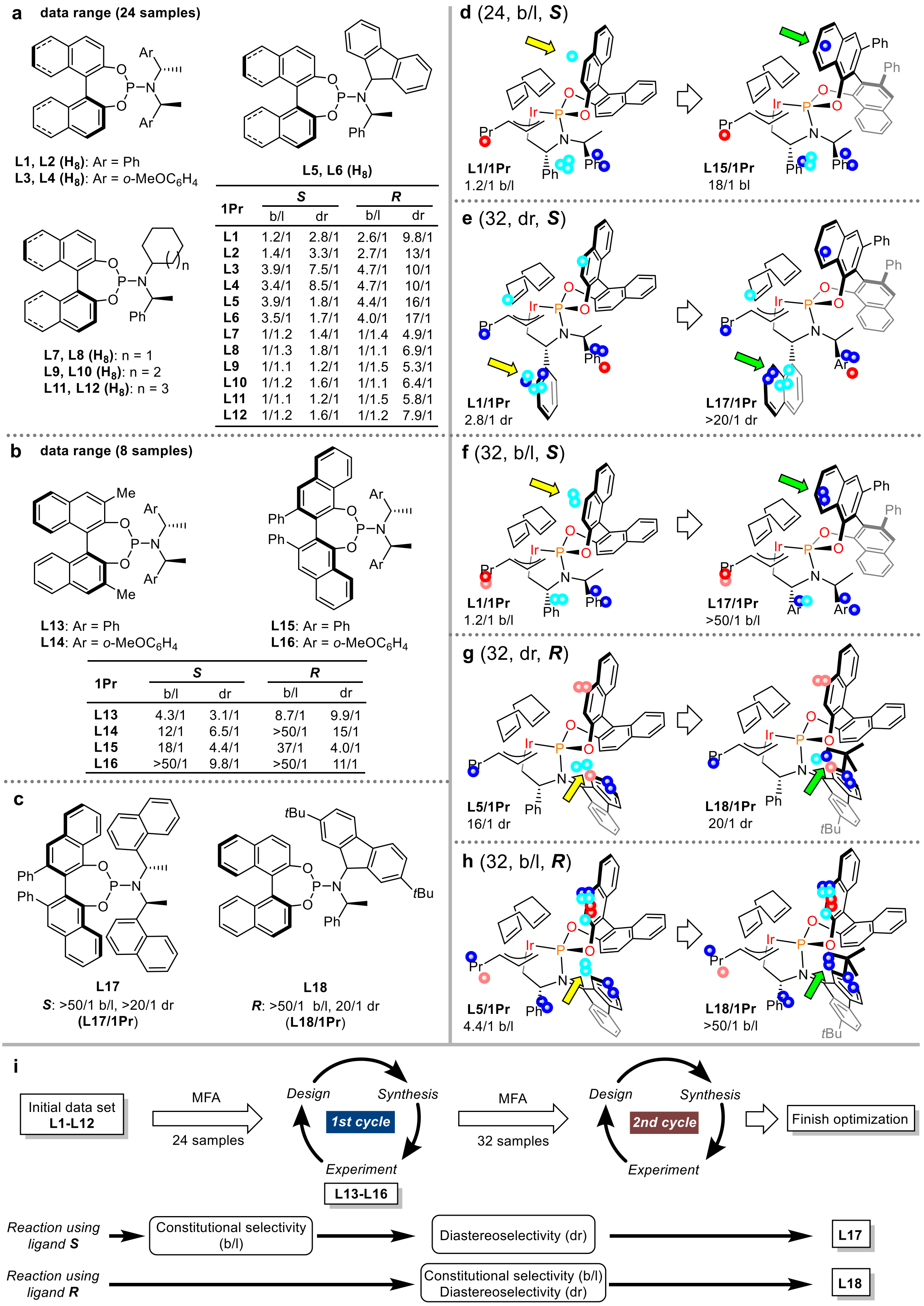
**Data-driven catalyst optimization.** When collecting training samples, mathematical indices are useful for selecting catalysts from a large number of candidate molecules as demonstrated by Denmark's group[4]. Typically, however, available catalysts are limited due to cost, synthetic accessibility, and catalytic activity. Thus, to construct the initial training datasets, phosphoramidite ligands for the Ir catalyst were selected based on their availability and catalytic activity (Fig. 2a). We employed **1Et** and **1Pr** as model substrates (Fig. 1c). Screening the reactions by combinations of 12 phosphoramidite ligands and two substrates afforded 4 sets of 24 target variables (b/l and dr using either the ***S*** or ***R*** ligand for the B catalyst, see Fig. 2a and see Supplementary Table S1–S4). We then performed MFA using the datasets (for details of the MFA, see Supplementary Information section 11). As the molecular fields, indicator fields (steric fields represented by the 0,1 vector as shown in Fig. 1d and see Supplementary fig. S1) were employed, which are often used for MFA in asymmetric catalysis[4,40,49-51]. It has been recognized that weak attractive interactions, such as dispersion forces, are sometimes important for asymmetric catalysis[52-55]. Because such weak non-covalent interactions are operative when atoms are present in sufficiently close proximity[52-55], positional information (3-dimensional coordinates) of atoms potentially functions as a descriptor to represent such attractive interactions. Indicator fields are calculated at each unit cell of a grid space. The unit cells that include the van der Waals radii of any atoms are counted as 1, or are otherwise counted as 0 (Fig. 1d and see Supplementary fig. S1). Therefore, while these descriptors mainly represent steric effects, the positional information also implicitly includes intermolecular weak interactions. Specifically, it is possible that the visualized information shown in Fig. 1d includes attractive dispersion interactions between Ir-π-allyl complexes and B-enolates. Thus, we expect that the indicator fields will be suitable for the MFA in this catalytic system. As target variables, logarithms of constitutional and diastereomeric ratios (b/l and dr) ($\Delta\Delta G^{\ddagger} = -RT\log$(b/l or dr)) were employed, which correspond to energy differences between the transition states that lead to each isomer (Curtin-Hammett principle[56]). By correlating the target variables and the molecular fields through LASSO[57] or Elastic Net[58] using the R package, glmnet[59] according to the reported procedure[51], four regression models were generated for predicting b/l and dr in the reactions using the ***S*** or ***R*** ligand.

Among the constructed regression models, first, we focused on the model dictating b/l ratios using the valine-derived ligand ***S*** for the boron catalyst, which produces (2*R*,3*R*)-**2** as the major isomer (1/1.2–5.0/1 b/l, 1/1.1–8.5/1 dr for the 24 reactions). From the coefficients of the regression model, important structural information was visualized on the Ir-π-allyl complex derived from **L1** and **1Pr** (hereafter referred to as **L1/1Pr**) affording 1.2/1 b/l (Fig. 2d). The blue and red points represent the structural information visualized based on the regression coefficients, corresponding to positive and negative coefficients, respectively—i.e., if molecular structures overlap with the blue/red points, selectivity increases/decreases (see the footnote of Fig. 1d for more details). A light blue point that did not overlap with intermediate **L1/1Pr** was identified between the 3- and 4-positions of the binaphthyl skeleton (yellow arrow), indicating that increasing the steric demand in this region should improve the b/l ratio. Thus, we employed (*S*)-3,3'-dimethyl BINOL and (*S*)-VANOL as the ligand skeletons, and synthesized ligands **L13** and **L15** (Fig. 2b). Similarly, **L14** and **L16** could be designed from the structure **L3/1Pr** (fig. S4). The intermediates **L13/1Pr**–**L16/1Pr** overlapped with the light blue points (e.g. **L15/1Pr**; Fig. 2d, the fused benzene ring of **L15** overlapping with the light blue point observed in **L1/1Pr** is indicated by the green arrow. See also Supplementary fig. S4). Indeed, the reactions using ligands **L13**–**L16** exhibited improved b/l ratios (4.3/1–>50/1) compared with that of **L1/1Pr** (1.2/1 b/l). In particular, the reactions using ligands **L15** and **L16** synthesized from (*S*)-VANOL exhibited a much higher b/l ratio than those in the initial training dataset (Figs. 2a and 2b, and see Supplementary Table S2).

While the b/l ratio increased after the optimization, dr had room for improvement (maximum 9.8/1 dr in the 32 samples using ligand ***S***). Thus, we again performed MFA on the data obtained from the 32 reactions, including those using ligands **L13**–**L16**. For the model from the dr data, a light blue point appeared near the Ph group at a phenethylamine moiety under the π-allyl group of **L1/1Pr** (yellow arrow in Fig. 2E). In the MFA of the b/l data from the 32 reactions, a light blue point again was identified around the 3- and 4-positions of the BINOL skeleton (Fig. 2f, yellow arrow). To increase both dr and the b/l ratio by superposing the light blue points near the Ph group on the amine moiety and the BINOL skeleton, we designed **L17** derived from (*S*)-VANOL bearing a naphthyl group on the amine moiety (Fig. 2c). The intermediate structure **L17/1Pr** overlapped with the blue points (green arrows in Figs. 2e and 2f, and fig. S5). To our delight, the reaction using **L17/1Pr** and ligand ***S*** exhibited excellent constitutional selectivity and diastereoselectivity (>50/1 b/l, >20/1 dr).

We next focused on the design of the Ir catalysts for the reactions using ligand ***R*** that produce the (2*S*,3*R*)-diastereomer as the major isomer. Simply changing the ligand for the B catalyst from ***S*** to ***R***, however, afforded (2*S*,3*R*)-**2Pr** in only 1.9/1 dr for the reaction using **L17** (see Supplementary Table S2). This result was again in contrast to Carreira's observation (Fig. 1a)[19], and suggested that enantiofacial controls of the Ir-π-allyl complex and B-enolate were not independent in our case. Thus, we screened reactions using the above-designed ligands **L13**–**L16** in addition to the initial training dataset (1/1.5–>50/1 b/l, 3.1/1–17/1 dr for the 32 reactions). Although **L14/1Pr** showed excellent b/l and high dr (>50/1 b/l, 15/1 dr), dr of **L5/1Pr** was higher (4.4/1 b/l, 16/1 dr). Therefore, we examined whether the b/l ratio could be further improved through data-driven optimization of **L5**. From the coefficients of the regression models, important structural information for b/l and dr was visualized on structure **L5/1Pr** (Figs. 2G and 2H). In both cases, light blue points appeared near the 2-position of the fluorene group (yellow arrow), indicating that introducing a substituent at the 2-position would improve both b/l and dr. According to this information, we designed **L18** bearing a *t*Bu group on the fluorene moiety (Fig. 2c). The structure **L18/1Pr** overlapped with the dark blue points (green arrows in Figs. 2g and 2h, and see Supplementary fig. S6). Gratifyingly, the reaction using **L18/1Pr** and ***R*** showed excellent constitutional selectivity with further improved diastereoselectivity (>50/1 b/l, 20/1 dr).

On the basis of the optimization studies, **L17** with ***S*** and **L18** with ***R*** are the optimum ligand combinations for the Ir/B dual catalysis to synthesize (2*R*,3*R*)-**2** and (2*S*,3*R*)-**2**, respectively (Fig. 2c). The ligands were identified through iterative MFA-based molecular design starting from easily accessible ligands (Fig. 2i). While we demonstrated the single design pathway, the pathway can be modified depending on the researcher’s intuition and starting training datasets. Another design pathway is shown in fig. S14. In addition, insights into the diastereomeric induction mechanisms can be obtained from the visualized structural information (see Supplementary figs. S19 and S20). While enantiomeric excess values in the initial training datasets were already high (90–99% ee in the 24 reactions using ***S*** and 92–99% ee in the 24 reactions using ***R***), we also performed MFA using the enantioselectivity data (see Supplementary figs. S26–S29). We predicted that designed ligands **L13**–**L18** would show excellent enantioselectivity over 97.8% (predicted $\Delta\Delta G^{\ddagger}$ > 2.9 kcal/mol, see Supplementary Table S12). Indeed, the reactions using those ligands furnished 98% ee or higher (see Supplementary Tables S1 and S2). Therefore, we would be able to design new ligands to improve enantioselectivity if necessary.



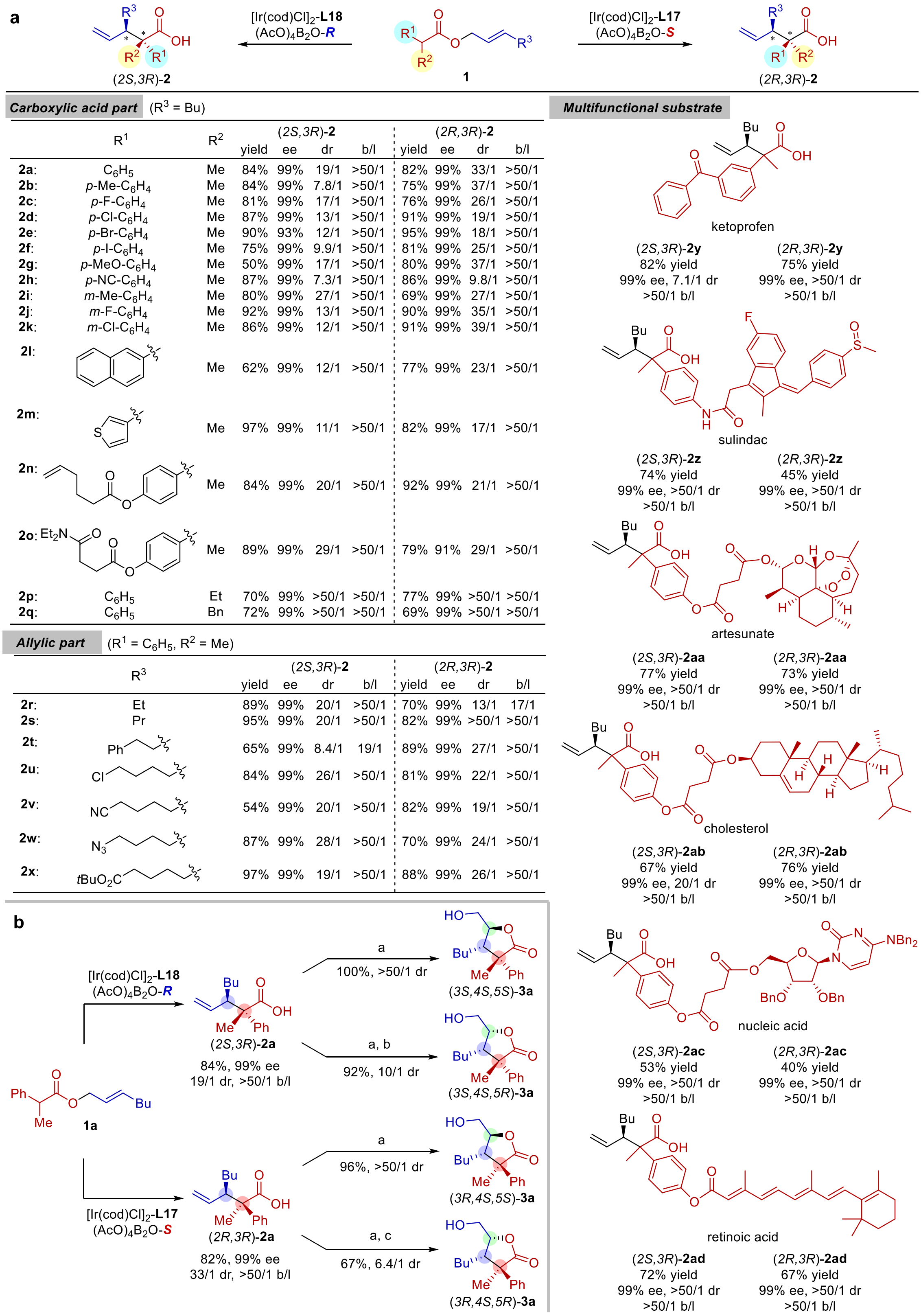
| 1Pr | S |  | R |  |
|---|---|---|---|---|
|  | b/l | dr | b/l | dr |
| **L1** | 1.2/1 | 2.8/1 | 2.6/1 | 9.8/1 |
| **L2** | 1.4/1 | 3.3/1 | 2.7/1 | 13/1 |
| **L3** | 3.9/1 | 7.5/1 | 4.7/1 | 10/1 |
| **L4** | 3.4/1 | 8.5/1 | 4.7/1 | 10/1 |
| **L5** | 3.9/1 | 1.8/1 | 4.4/1 | 16/1 |
| **L6** | 3.5/1 | 1.7/1 | 4.0/1 | 17/1 |
| **L7** | 1/1.2 | 1.4/1 | 1/1.4 | 4.9/1 |
| **L8** | 1/1.3 | 1.8/1 | 1/1.1 | 6.9/1 |
| **L9** | 1/1.1 | 1.2/1 | 1/1.5 | 5.3/1 |
| **L10** | 1/1.2 | 1.6/1 | 1/1.1 | 6.4/1 |
| **L11** | 1/1.1 | 1.2/1 | 1/1.5 | 5.8/1 |
| **L12** | 1/1.2 | 1.6/1 | 1/1.2 | 7.9/1 |

| 1Pr | S |  | R |  |
|---|---|---|---|---|
|  | b/l | dr | b/l | dr |
| **L13** | 4.3/1 | 3.1/1 | 8.7/1 | 9.9/1 |
| **L14** | 12/1 | 6.5/1 | >50/1 | 15/1 |
| **L15** | 18/1 | 4.4/1 | 37/1 | 4.0/1 |
| **L16** | >50/1 | 9.8/1 | >50/1 | 11/1 |

**Fig. 2 | Dataset and molecular design. a**, An initial dataset. The phosphoramidite ligands for the Ir catalysts derived from BINOL or $H_8$BINOL (**$H_8$**) were employed for the reactions using substrates **1Et** and **1Pr**. The reaction conditions were identical to Fig. 1c. **b**–**c**, Ligands designed by the MFA. **d**–**h**, Important structural information visualized on the Ir-π-allyl intermediates and molecular design based on the structural information.

While the guidelines were visualized on the CPK models of the intermediate structures as shown in Fig. 1d, we show the structural information and the intermediate structures drawn by ChemDraw for clarity. The structural information was extracted by LASSO. The number in parenthesis is the number of the reactions used for the MFA. **Ligand/substrate** denotes the Ir-π-allyl complex structure composed of the ligand and substrate or the reaction using the ligand and substrate. The b/l or dr determined by the experiments are shown. **i**, Overall design flow.

**Substrate scope and synthetic applications.** Next, we investigated the substrate scope of these asymmetric dual catalyst systems (Fig. 3a). A range of allyl esters bearing alkyl, halogens, electron-withdrawing groups, electron-donating substituents, and a heteroatom at the aromatic ring ($R^1$) provided products (2*S*,3*R*)-**2** and (2*R*,3*R*)**-2** with high enantio-, diastereo-, and constitutional selectivity (**2a**–**2m**). The allyl groups were introduced chemoselectively at the α-position of the carboxylic acid function in the presence of an ester (**2n**) or amide function (**2o**) bearing intrinsically more acidic α-C–H bonds. This was due to the chemoselective activation of carboxylic acids by the B catalyst[34]. Substituent $R^2$ can be replaced by alkyl groups other than a methyl group while maintaining high stereoselectivity (**2p**, **2q**). Various functional groups at the allyl moiety were also tolerated (**2r**–**2x**), including alkyl chloride (**2u**) and alkyl azide (**2w**). Due to the high functional group compatibility, the reaction was applied to late-stage modification of multifunctional molecules, including substrates derived from anti-inflammatory drugs (ketoprofen [**2y**] and sulindac [**2z**]), an anti-malarial drug (artesunate [**2aa**]), a cholesterol (**2ab**), a nucleic acid (**2ac**) and a retinoic acid (**2ad**). In all entries, the reactions proceeded with catalyst-controlled stereoselectivity and constitutional selectivity. To demonstrate synthetic utility, we examined stereodivergent construction of three contiguous stereocenters (Fig. 3b). Thus, osmium-catalyzed alkene dihydroxylation of (*2S*,*3R*)-**2a** and (*2R*,*3R*)-**2a** afforded diastereomerically pure lactones (*3S*,*4S*,*5S*)-**3a** and (*3R*,*4S*,*5S*)-**3a** in high yield, respectively. Inversion of the stereochemistry at C5 proceeded through *O*-mesylation followed by hydrolysis, affording (*3S*,*4S*,*5R*)-**3a** and (*3R*,*4S*,*5R*)-**3a**, respectively. This result showcases that 8 stereoisomers (including enantiomers) can be divergently synthesized from a simple starting material by combining the Ir/B dual asymmetric catalysis with traditional methods.



**Carboxylic acid part** ($R^3$ = Bu)

| | $R^1$ | $R^2$ | (2*S*,3*R*)-2 yield | ee | dr | b/l | (2*R*,3*R*)-2 yield | ee | dr | b/l |
|---|---|---|---|---|---|---|---|---|---|---|
| **2a**: | $C_6H_5$ | Me | 84% | 99% | 19/1 | >50/1 | 82% | 99% | 33/1 | >50/1 |
| **2b**: | *p*-Me-$C_6H_4$ | Me | 84% | 99% | 7.8/1 | >50/1 | 75% | 99% | 37/1 | >50/1 |
| **2c**: | *p*-F-$C_6H_4$ | Me | 81% | 99% | 17/1 | >50/1 | 76% | 99% | 26/1 | >50/1 |
| **2d**: | *p*-Cl-$C_6H_4$ | Me | 87% | 99% | 13/1 | >50/1 | 91% | 99% | 19/1 | >50/1 |
| **2e**: | *p*-Br-$C_6H_4$ | Me | 90% | 93% | 12/1 | >50/1 | 95% | 99% | 18/1 | >50/1 |
| **2f**: | *p*-I-$C_6H_4$ | Me | 75% | 99% | 9.9/1 | >50/1 | 81% | 99% | 25/1 | >50/1 |
| **2g**: | *p*-MeO-$C_6H_4$ | Me | 50% | 99% | 17/1 | >50/1 | 80% | 99% | 37/1 | >50/1 |
| **2h**: | *p*-NC-$C_6H_4$ | Me | 87% | 99% | 7.3/1 | >50/1 | 86% | 99% | 9.8/1 | >50/1 |
| **2i**: | *m*-Me-$C_6H_4$ | Me | 80% | 99% | 27/1 | >50/1 | 69% | 99% | 27/1 | >50/1 |
| **2j**: | *m*-F-$C_6H_4$ | Me | 92% | 99% | 13/1 | >50/1 | 90% | 99% | 35/1 | >50/1 |
| **2k**: | *m*-Cl-$C_6H_4$ | Me | 86% | 99% | 12/1 | >50/1 | 91% | 99% | 39/1 | >50/1 |
| **2l**: | | Me | 62% | 99% | 12/1 | >50/1 | 77% | 99% | 23/1 | >50/1 |
| **2m**: | | Me | 97% | 99% | 11/1 | >50/1 | 82% | 99% | 17/1 | >50/1 |
| **2n**: | | Me | 84% | 99% | 20/1 | >50/1 | 92% | 99% | 21/1 | >50/1 |
| **2o**: | | Me | 89% | 99% | 29/1 | >50/1 | 79% | 91% | 29/1 | >50/1 |
| **2p**: | $C_6H_5$ | Et | 70% | 99% | >50/1 | >50/1 | 77% | 99% | >50/1 | >50/1 |
| **2q**: | $C_6H_5$ | Bn | 72% | 99% | >50/1 | >50/1 | 69% | 99% | >50/1 | >50/1 |

**Allylic part** ($R^1$ = $C_6H_5$, $R^2$ = Me)

| | $R^3$ | (2*S*,3*R*)-2 yield | ee | dr | b/l | (2*R*,3*R*)-2 yield | ee | dr | b/l |
|---|---|---|---|---|---|---|---|---|---|
| **2r**: | Et | 89% | 99% | 20/1 | >50/1 | 70% | 99% | 13/1 | 17/1 |
| **2s**: | Pr | 95% | 99% | 20/1 | >50/1 | 82% | 99% | >50/1 | >50/1 |
| **2t**: | | 65% | 99% | 8.4/1 | 19/1 | 89% | 99% | 27/1 | >50/1 |
| **2u**: | | 84% | 99% | 26/1 | >50/1 | 81% | 99% | 22/1 | >50/1 |
| **2v**: | | 54% | 99% | 20/1 | >50/1 | 82% | 99% | 19/1 | >50/1 |
| **2w**: | | 87% | 99% | 28/1 | >50/1 | 70% | 99% | 24/1 | >50/1 |
| **2x**: | | 97% | 99% | 19/1 | >50/1 | 88% | 99% | 26/1 | >50/1 |

**Fig. 3 | Substrate scope of the stereodivergent asymmetric dual catalysis and transformations of the products. a**, For the reaction conditions, see the footnote of Fig. 1C, except for reactions of **1z**, **1aa**, **1ac** and **1ad** [cat. Ir* ([Ir(cod)Cl]$_2$ (5 mol%)–**L17** or **L18** (10 mol%)) and cat. B* ((AcO)$_4$B$_2$O (20 mol%)–***S*** or ***R*** (40 mol%)). **b**, Stereodivergent construction of three contiguous stereocenters. Reaction conditions: (a) **2a** (1.0 equiv), OsO$_4$ (10 mol%), and *N*-methylmorphiline (1.5 equiv) were mixed in *t*BuOH/H$_2$O for 12 h at rt. (b) (3*S*,4*S*,5*S*)-**3**, MsCl (2.4 equiv), and triethylamine (2.0 equiv) mixed in DCM for 3 h at rt; mesylated product (1.0 equiv) and KOH (3.0 equiv) mixed in THF/H$_2$O for 3 h at rt; (c) (3*R*,4*S*,5*S*)-**3**, MsCl (2.4 equiv), and triethylamine (2.0 equiv) mixed in DCM for 3 h at rt; mesylated product (1.0 equiv) and tetramethylammonium hydroxide pentahydrate (3.0 equiv) were mixed in THF for 3 h at 0 °C.

## Conclusion

In conclusion, we successfully improved the four reaction outcomes (constitutional selectivity and diastereoselectivity for both diastereomers) while retaining excellent enantioselectivity, in stereodivergent asymmetric Ir/B dual hybrid catalysis. Our method involved MFA-based data-driven catalyst optimization with a relatively small dataset, based on the Ir-π-allyl complex intermediate structures without knowing the detailed structures of the reaction partner (B-enolate). The concept demonstrated herein will enable efficient optimization of various stereodivergent asymmetric hybrid catalysis. Our analysis employed simple steric descriptors, the indicator fields, and standard machine learning techniques, LASSO[57] and Elastic Net[58]. Further examinations of molecular fields and machine learning techniques will expand the scope of this analysis and accelerate the development of data science in molecular catalysis.

## References

1. Sanchez-Lengeling, B., & Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* ***361***, 360–365 (2018).
2. Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. Predicting reaction performance in C-N cross-coupling using machine learning. *Science* ***360***, 186–190 (2018).
3. Granda, J. M., Donina, L., Dragone, V., Long, D.-L., & Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* ***559***, 377–381 (2018).
4. Zahrt, A. F., Henle, J. J., Rose, B. T., Wang, Y., Darrow, W. T. & Denmark, S. E. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* ***363***, eaau5631 (2019).
5. Chu, Y., Heyndrickx, W., Occhipinti, G., Jensen, V. R. & Alsberg, B. K. An evolutionary algorithm for de novo optimization of functional transition metal compounds. *J. Am. Chem. Soc.* ***134***, 8885–8895 (2012).
6. Burello, E., Farrusseng, D. & Rothenberg, G. Combinatorial explosion in homogeneous catalysis: Screening 60,000 cross-coupling reactions. *Adv. Synth. Catal.* ***346***, 1844–1853 (2004).
7. Gao, H., Struble, T. J., Coley, C. W., Wang, Y., Green, W. H. & Jensen, K. F. Using machine learning to predict suitable conditions for organic reactions. *ACS. Cent. Sci.* ***4***, 1465–1476 (2018).
8. Singh, S. Pareek, M., Changotra, A., Banerjee, P., Bhaskararao, B., Balamurugan, P. & Sunoj, R. B. A unified machine-learning protocol for asymmetric catalysis as a proof of concept demonstration using asymmetric hydrogenation. *Proc. Natl. Acad. Sci. USA* ***117***, 1339–1345 (2020).
9. Williams, A. *Free Energy Relationships in Organic and Bio-Organic Chemistry* (Royal Society of Chemistry, 2003).
10. Sigman, M. S., Harper, K. C., Bess, E. N. & Milo, A. The development of multidimensional analysis tools for asymmetric catalysis and beyond. *Acc. Chem. Res.* ***49***, 1292–1301 (2016).

11. Santiago, C. B., Guo, J.-Y. & Sigman, M. S. Predictive and mechanistic multivariate linear regression models for reaction development. *Chem. Sci. 9*, 2398–2412 (2018).

12. Zhao, S., Gensch, T., Murray, B., Niemeyer, Z. L., Sigman, M. S. & Biscoe, M. R. Enantiodivergent Pd-catalyzed C–C bond formation enabled through ligand parameterization *Science* ***362***, 670–674 (2018).

13. Reid, J. P. & Sigman, M. S. Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature* ***571***, 343–348 (2019).

14. Fujita, T. & Winkler, D. A. Understanding the roles of the "two QSARs". *J. Chem. Inf. Model.* ***56***, 269–274 (2016).

15. Zahrt, A. F., Athavale, S. V. & Denmark, S. E. Quantitative structure–selectivity relationships in enantioselective catalysis: Past, present, and future. *Chem. Rev.* ***120***, 1620–1689 (2020).

16. Toyao, T., Maeno, Z., Takakusagi, S., Kamachi, T., Takigawa, I. & Shimizu, K. Machine learning for catalysis informatics: Recent applications and prospects. *ACS Catal*. ***10***, 2260-2297 (2020).

17. Foscato, M. & Jensen, V. R. Automated in silico design of homogeneous catalysts. *ACS Catal.* ***10***, 2354-2377 (2020).

18. Jacobsen, E. N., Pfaltz, A. & Yamamoto, H. Eds., *Comprehensive Asymmetric Catalysis: Vol. I-III, Suppl. I-II* (Springer, New York, 1999).

19. Krautwald, S., Sarlah, D., Schafroth, M. A. & Carreira, E. M. Enantio- and diastereodivergent dual catalysis: $\alpha$-allylation of branched aldehydes. *Science* ***340***, 1065–1068 (2013).

20. Krautwald, S. & Carreira, E. M. Stereodivergence in asymmetric catalysis. *J. Am. Chem. Soc.* ***139***, 5627−5639 (2017).

21. Huo, X., He, R., Zhang, X., & Zhang, W. An Ir/Zn Dual Catalysis for Enantio- and Diastereodivergent α-Allylation of α-Hydroxyketones. *J. Am. Chem. Soc.* ***138***, 11093–11096 (2016).

22. He, R., Liu, P., Huo, X., & Zhang, W. Ir/Zn Dual Catalysis: Enantioselective and Diastereodivergent α-Allylation of Unprotected α-Hydroxy Indanones. *Org. Lett.* ***19***, 5513–5516 (2017).

23. Liu, X.-J., Jin, S., Zhang, W.-Y., Liu, Q.-Q., Zheng, C. & You, S.-L. Sequence-Dependent Stereodivergent Allylic Alkylation/Fluorination of Acyclic Ketones. *Angew. Chem., Int. Ed.* ***59***, 2039–2043 (2020).

24. Jiang, X., Beiger, J. J. & Hartwig, J. F. Stereodivergent Allylic Substitutions with Aryl Acetic Acid Esters by Synergistic Iridium and Lewis Base Catalysis. *J. Am. Chem. Soc.* ***139***, 87–90 (2017).

25. Wei, L., Zhu, Q., Xu, S.-M., Chang, X. & Wang, C.-J. Stereodivergent Synthesis of α,α-Disubstituted α-Amino Acids via Synergistic Cu/Ir Catalysis. *J. Am. Chem. Soc.* ***140***, 1508–1513 (2018).

26. Huo, X., Zhang, J., Fu, J., He, R. & Zhang, W. Ir/Cu Dual Catalysis: Enantio- and Diastereodivergent Access to α,α-Disubstituted α-Amino Acids Bearing Vicinal Stereocenters. *J. Am. Chem. Soc.* ***140***, 2080–2084 (2018).

27. Zhang, J., Huo, X., Li, B., Chen, Z., Zou, Y., Sun, Z. & Zhang, W. Enantioselective and Diastereodivergent Access to a-Substituted a-Amino Acids via Dual Iridium and Copper Catalysis. *Adv. Synth. Catal.* ***361***, 1130–1139 (2019).

28. He, Z. T., Jiang, X. & Hartwig, J. F. Stereodivergent Construction of Tertiary Fluorides in Vicinal Stereogenic Pairs by Allylic Substitution with Iridium and Copper Catalysts. *J. Am. Chem. Soc.* ***141***, 13066–13073 (2019).

29. Wu, H.-M., Zhang, Z., Xiao, F., Wei, L., Dong, X.-Q. & Wang, C.-J. Stereodivergent Synthesis of α-Quaternary Serine and Cysteine Derivatives Containing Two Contiguous Stereogenic Centers via Synergistic Cu/Ir Catalysis. *Org. Lett.* ***22***, 4852–4857 (2020).

30. Jiang, X., Boehm, P. & Hartwig, J. F. Stereodivergent Allylation of Azaaryl Acetamides and Acetates by Synergistic Iridium and Copper Catalysis. *J. Am. Chem. Soc.* ***140***, 1239–1242 (2018).

31. Ohmura, T. & Hartwig, J. F. Regio- and Enantioselective Allylic Amination of Achiral Allylic Esters Catalyzed by an Iridium-Phosphoramidite Complex. *J. Am. Chem. Soc.* ***124***, 15164–15165 (2002).

32. Bartels, B., Garcia-Yebra, C. & Helmchen, G. Asymmetric $Ir^{I}$-Catalysed Allylic Alkylation of Monosubstituted Allylic Acetates with Phosphorus Amidites as Ligands. *Eur. J. Org. Chem.* 1097–1103 (2003).

33. Cheng, Q., Tu, H.-F., Zheng, C., Qu, J.-P., Helmchen, G. & You, S.-L. Iridium-catalyzed asymmetric allylic substitution reactions. *Chem. Rev.* ***119***, 1855−1969 (2019).

34. Fujita, T., Yamamoto, T., Morita, Y., Chen, H., Shimizu, Y. & Kanai, M. Chemo- and enantioselective Pd/B hybrid catalysis for the construction of acyclic quaternary carbons: Migratory allylation of *O*-allyl esters to *α*-*C*-allyl carboxylic acids. *J. Am. Chem. Soc.* ***140***, 5899–5903 (2018).

35. Cruz, F. A. & Dong, V. M. Stereodivergent coupling of aldehydes and alkynes via synergistic catalysis using Rh and Jacobsen's amine. *J. Am. Chem. Soc.* ***139***, 1029–1032 (2017).

36. Pierrot, D. & Marek, I. Synthesis of enantioenriched vicinal tertiary and quaternary carbon stereogenic centers within an acyclic chain. *Angew. Chem., Int. Ed.* ***59***, 36–49 (2020).

37. Cramer, R. D., Patterson, D. E. & Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* ***110***, 5959–5967 (1988).

38. Lipkowitz, K. B. & Pradhan, M. Computational studies of chiral catalysts: A comparative molecular field analysis of an asymmetric Diels−Alder reaction with catalysts containing bisoxazoline or phosphinooxazoline Ligands. *J. Org. Chem.* ***68***, 4648–4656 (2003).

39. Kozlowski, M. C., Dixon, S. L., Panda, M. & Lauri, G. Quantum mechanical models correlating structure with selectivity: Predicting the enantioselectivity of *β*-Amino Alcohol Catalysts in Aldehyde Alkylation. *J. Am. Chem. Soc.* ***125***, 6614–6615 (2003).

40. Yamaguchi, S. & Sodeoka, M. Molecular field analysis using intermediates in enantio-determining steps can extract information for data-driven molecular design in asymmetric catalysis. *Bull. Chem. Soc. Jpn.* ***92***, 1701–1706 (2019).

41. Hansch, C., Maloney, P. P., Fujita, T., & Muir. R. M. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature* ***194***, 178–180 (1962).

42. Hansch, C. & Fujita, T. p-σ-π analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* ***86***, 1616–1626 (1964).

43. Hansch, C., Leo, A. & Hoekman, D. H. Exploring QSAR, Fundamentals and application in chemistry and biology; American Chemical Society: Washington, DC, USA (1995).

44. Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y. C., Todeschini, R., Consonni, V., Kuz'min, V. E., Cramer, R., Benigni, R., Yang, C., Rathman, J., Terfloth, L., Gasteiger, J., Richard, A. & Tropsha, A. QSAR modeling: Where have you been? Where are you going to? *J. Med. Chem.* ***57***, 4977–5010 (2014).

45. Hartwig, J. F. & Stanley, L. M. Mechanistically driven development of iridium catalysts for asymmetric allylic substitution. *Acc. Chem. Res.* ***43***, 1461–1475 (2010).

46. Bhaskararao, B. & Sunoj, R. B. Origin of Stereodivergence in Cooperative Asymmetric Catalysis with Simultaneous Involvement of Two Chiral Catalysts. *J. Am. Chem. Soc.* ***137***, 15712–15722 (2015).

47. Bhaskararao, B. & Sunoj, R. B. Asymmetric Dual Chiral Catalysis using Iridium Phosphoramidites and Diarylprolinol Silyl Ethers: Insights into Stereodivergence. *ACS Catal.* ***7***, 6675–6685 (2017).

48. Changotra, A., Bhaskararao, B., Hadad, C. M. & Sunoj, R. B. Insights on Absolute and Relative Stereocontrol in Stereodivergent Cooperative Catalysis. *J. Am. Chem. Soc.* ***142***, 9612–9624 (2020).

49. Melville, J. L., Lovelock, K. R. J., Wilson, C., Allbutt, B., Burke, E. K., Lygo, B. & Hirst, J. D. Exploring phase-transfer catalysis with molecular dynamics and 3D/4D quantitative structure−selectivity relationships. *J. Chem. Inf. Model.* ***45***, 971-981 (2005).

50. Denmark, S. E., Gould, N. D. & Wolf, L. M. A systematic investigation of quaternary ammonium ions as asymmetric phase-transfer catalysts. Application of quantitative structure activity/selectivity relationships. *J. Org. Chem.* ***76***, 4337-4357 (2011).

51. Yamaguchi, S., Nishimura, T., Hibe, Y., Nagai, M., Sato, H. & Johnston, I. Regularized regression analysis of digitized molecular structures in organic reactions for quantification of steric effects. *J. Comp. Chem.* ***38***, 1825–1833 (2017).

52. Wagner, J. P. & Schreiner, P. R. London dispersion in molecular chemistry— Reconsidering steric effects. *Angew. Chem. Int. Ed.* ***54***, 12274–12296 (2015).

53. Neel, A. J.; Hilton, M. J.; Sigman, M. S.; Toste, F. D. Exploiting non-covalent π interactions for catalyst design. *Nature* ***543***, 637–646 (2017).

54. Lu, G., Liu, R. Y., Yang, Y., Fang, C., Lambrecht, D. S., Buchwald, S. L. & Liu, P. Ligand–substrate dispersion facilitates the copper-catalyzed hydroamination of unactivated olefins. *J. Am. Chem. Soc.* ***139***, 16548−16555 (2017).

55. Xi, Y. Su, B., Qi, X. Pedram, S., Liu, P. & Hartwig, J. F. Application of trimethylgermanyl-substituted bisphosphine ligands with enhanced dispersion interactions to copper-catalyzed hydroboration of disubstituted alkenes. *J. Am. Chem. Soc.* **142**, 18213−18222 (2020).

56. Seeman, J. I. Effect of conformational change on reactivity in organic chemistry. Evaluations, application, and extensions of Curtin–Hammett/Winstein–Holness Kinetics". *Chem. Rev.* ***83***, 83–134 (1983).

57. Tibshirani, R. Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. B* ***58***, 267–288 (1996).

58. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* ***67***, 301–320 (2005).

59. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* ***33***, 1–22 (2010).

## Methods

**General procedure for Asymmetric Ir/B dual catalysis.** A flame-dried 10 mL test tube A, equipped with a magnetic stirring bar, was charged with $[Ir(cod)Cl]_2$ (5.0 mg, 0.0075 mmol, 0.025 equiv), ligand **L17** or **L18** (0.015 mmol, 0.05 equiv), LiCl (12.8 mg, 0.3 mmol, 1 equiv), DBU (67.2 μL, 0.45 mmol, 1.5 equiv), and anhydrous THF (125 μL). Another flame-dried 10 mL test tube B, equipped with a magnetic stirring bar, was charged with $(AcO)_4B_2O$ (8.2 mg, 0.030 mmol, 0.1 equiv), (*R*)- or (*S*)-((2,3,5,6-tetrafluoro-4-methoxyphenyl)sulfonyl)-valine (***R*** or ***S***, 21.6 mg, 0.060 mmol, 0.2 equiv), and anhydrous THF (125 μL). After stirring the solution in test tube A for 1 h at 50 °C and stirring the solution in test tube B for 1 h at room temperature, the solution in test tube B and allyl ester **1** (0.30 mmol, 1 equiv) were added sequentially to the test tube A. The reaction mixture was stirred for 12 h at 50 °C under argon atmosphere. The reaction was quenched with aq. HCl (1.0 M) and products were extracted with EtOAc. The organic layer was washed with aqueous HCl (1.0 M), dried over $Na_2SO_4$, filtered and concentrated under reduced pressure to afford the crude product. Diastereomeric ratio (dr) and branch/linear (b/l) selectivity were determined by $^1$H NMR analysis of crude mixture. The crude product was purified by column chromatography (hexane/EtOAc = 5/1).

**Calculations of the molecular fields.** The protocol for calculations of molecular fields is as follows (see Supplementary fig. S1a). (I) A set of the Ir-π-allyl intermediates was optimized using the DFT method at the B3LYP/LANL2DZ (Ir) and 6-31G(d) level. (II) The coordinates of the set of the molecules obtained in step I were aligned. For alignment, first, we defined an xy plane based on the mean plane of the allyl group of **L1/1Et** as shown in fig. S1c. The central carbon atom of the allyl group was set as the origin. Then, alignment was performed through the least squares method by minimizing the distances between the allyl groups (the 7 atoms highlighted by red shown in fig. S1b) of **L1/1Et** and other intermediates. (III) The structures were placed in a grid space. The unit cell size is 1 Å per side. We used the molecular structures around the reaction center for calculations of the molecular fields instead of the use of the whole molecular structures to reduce dimensions of descriptors and avoid overfitting. The size of the grid space, which is centered at the origin, is 10 × 12 × 6 Å$^3$. Each unit cell is regarded as an element of the descriptor vectors. The unit cells that included Bondi van der Waals radii of C (1.70 Å), H (1.20 Å), O (1.52 Å) atoms were counted as 1, or were otherwise counted as 0. Columns of all 0 and all 1 were removed to give the descriptor matrix.

**LASSO and Elastic Net regression.** Logarithms of constitutional selectivity (b/l) and diastereoselectivity (dr) were employed as target variables (see Supplementary Tables S3 and S4), which correspond to energy differences between the transition states that lead to each isomer ($\Delta\Delta G^{\ddagger} = -RT\log(\text{b/l or dr})$). $R$ is the gas constant and $T$ is the temperature of the reactions, 323.15 K. The molecular fields and $\Delta\Delta G^{\ddagger}$ values (kcal/mol) were correlated using LASSO (Least Absolute Shrinkage and Selection Operator) or Elastic Net to generate the regression models. The LASSO and Elastic Net regressions were performed using the R package, glmnet.

$$E(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}\left\{(1-\alpha)\beta_j^2 + \alpha|\beta_j|\right\}$$

By minimizing the above loss function, we can estimate coefficients $\beta = (\beta_1,\beta_2,\ldots,\beta_j,\ldots,\beta_p)^T$ while simultaneously assigning unimportant coefficients for reaction outcomes to be 0 (In this study, $y_i$ is $\Delta\Delta G^{\ddagger}$. $\mathrm{x}_i = (x_{i1}, x_{i2},\ldots,x_{ij},\ldots,x_{ip})^T$ is the indicator field. $n$ and $p$ denote the number of samples and descriptors, respectively.). If $\alpha$ is 1, this method corresponds to LASSO regression. If $\alpha$ is $0 < \alpha < 1$, this method corresponds to Elastic Net regression. In all cases in this study, we selected values of the hyper parameter□$\lambda$ that minimized the mean squared error calculated from predicted values obtained from leave-one-out cross-validation by using glmnet. Among descriptors that are correlated, LASSO will select one. In contrast, Elastic Net can extract multiple correlated descriptors. We performed Elastic Net regression if structural information that led to molecular design could not be found through LASSO regression. For Elastic Net, we varied the parameter $\alpha$ from 0.1 to 0.9 in steps of 0.1, and for each choice of $\alpha$ we selected the parameter $\lambda$ according to the procedure described above using glmnet. Among the models in which the important structural information that led to molecular design was included, we employed the model showing the highest $q^2$ (coefficient of determination calculated from predicted values of leave-one-out cross-validation [LOOCV]). Coefficients of determination calculated from the resulting regression models ($R^2$) and $q^2$ are shown in fig. S2 (24 reactions) and fig. S3 (32 reactions) along with plots of the measured and predicted values. The numbers of all descriptors and extracted descriptors are also shown in figs S2 and S3. The measured and predicted values are summarized in Table S3 and S4. We also performed 4-fold CV and y-randomization. The analysis was repeated 100 times for 4-fold CV and 50 times for y-randomization. The average values of the coefficients of determination are shown in fig. S2 and S3 ($Q^2$ for 4-fold CV and $R^2_{\text{yradom}}$ for y-randomization). In all cases, $Q^2$ showed good values over 0.6, indicating the models are robust. Low values of $R^2_{\text{yrandom}}$ close to 0 indicate the probability of chance correlation is low.

**Molecular design.** The workflow for the molecular design in this study is as follows.

(I) MFA using the sets of selectivity data and the intermediate structures is performed and the quality of the resulting regression models is checked. In this study, we have successfully designed the molecules based on structural information visualized by the MFA. All the regression models showed $R^2$, $q^2$, $Q^2 > 0.6$ and $R^2_{\text{y-random}} <$ 0.2. Thus, we tentatively employ these metrics ($R^2$, $q^2$, $Q^2 > 0.6$ and $R^2_{\text{y-random}} < 0.2$) to determine whether the models are used for the design.

(II) Extracted structural information is visualized on the intermediate structures and all the intermediate structures in the training samples and visualized structural information are thoroughly compared. By considering synthetic accessibility, substituents are introduced to the intermediate structures to overlap light blue points as shown in the text. For the design, we basically employ the structural information that fulfills $|r| > 0.3$ and shows the same sign with correlation coefficient $r$ (i.e. the structural information corresponding to positive regression coefficients should show a positive value of correlation coefficient $r$ with the target variables). If predicted $\Delta\Delta G^{\ddagger}$ values of the designed molecules are higher than those of the template molecules, the designed ligands are synthesized and the reactions using the ligands are examined.

(III) If the selectivity values are not satisfactory, MFA using all the sets of target variables including those in the reactions with the designed ligands is again performed. This workflow is repeated until designed molecules show high selectivity.

## Data availability

All data supporting the findings of this study, including experimental procedures and compound characterization, NMR, and HPLC are available within the Article and its Supplementary Information. Input data for data analysis is found in “data” folder in Supplementary Information. Important structural information with all the 36 Ir-π-allyl intermediate structures (xyz files) obtained by running our scripts is found in “output” folder of Supplementary Information. All the regression coefficient values including standardized regression coefficients in the regression models used for the molecular design along with correlation coefficient $r$ and coordinates of unit cell centers are summarized as excel files found in parameters folder of Supplementary Information.

## Code availability

Scripts for the MFA are available at https://github.com/sh-yamaguchi/MFA and the brief summary about how to use the scripts are described in Supplementary Information.

## Acknowledgement

We thank Yuki Nishioka for conducting preliminary experiments. This work was supported by JSPS KAKENHI Grant Numbers JP17H06441 (M.K.), JP20H04831 (S.Y.), JP20H04797 (Y.S.) (Hybrid Catalysis), JP20H05843 (H.M.) (Dynamic Exciton) and 18H05969 (H.M.). The DFT calculations were performed on the RIKEN HOKUSAI supercomputer system.

**Author contributions**: S.Y., Y.M., Y.S., H.M., and M.K. conceived and designed the project. H.C., Y.M., H.N., Z.Z., and H.M. carried out the experiments. S.Y. carried out data-driven molecular design. S.Y., H.M., and M.K. wrote the manuscript. All authors analyzed the data, discussed the results, and proofread the manuscript.

**Competing interests**: The authors declare no competing interests.

**Correspondence and requests for materials** should be addressed to S.Y., H.M., and M.K.