nature synthesis

Article

https://doi.org/10.1038/s44160-022-00233-y

Enantioselectivity prediction of pallada-electrocatalysed C–H activation using transition state knowledge in machine learning

Received: 28 March 2022

Accepted: 15 December 2022

Published online: 30 January 2023

Check for updates

Li-Cheng Xu[1], Johanna Frey[2], Xiaoyan Hou[2], Shuo-Qing Zhang[1], Yan-Yu Li[1], João C. A. Oliveira[2], Shu-Wen Li[1], Lutz Ackermann[2,3] ✉ & Xin Hong[1,4,5] ✉

Enantioselectivity prediction in asymmetric catalysis has been a long-standing challenge in synthetic chemistry because of the high-dimensional nature of the structure–enantioselectivity relationship. A lack of understanding of the synthetic space results in laborious and time-consuming efforts in the discovery of asymmetric reactions, even if the same transformation has already been optimized on model substrates. Here we present a data-driven workflow to achieve a holistic enantioselectivity prediction of asymmetric pallada-electrocatalysed C–H activation by implementing transition state knowledge in machine learning. The vectorization of transition state knowledge allowed for an excellent description and extrapolation of the machine learning model, and enabled the quantitative evaluation of 846,720 possibilities. Model interpretation revealed the non-intuitive olefin effect on the enantioselectivity determination. Subsequent density functional theory calculations unravelled mechanistic knowledge that the rate-determining step depends on the olefin reactivity in the insertion step. Therefore, the olefin insertion step can be involved in the overall enantioselectivity determination. These results highlight the complementary features of knowledge-based machine learning with an interpretation-driven mechanistic study, which provides the opportunity to harness widely existing catalysis screening data and transition state models in molecular synthesis.

As a key to enable chemical science, asymmetric catalysis plays a pivotal role in molecular synthesis[1,2]. Over the past few decades, intensive research in asymmetric catalysis reshaped this field, which serves and steers the escalating demand for chiral compounds from society[3,4]. The classic strategy to design and improve asymmetric catalysis is the mechanism-based approach (Fig. 1a)[5,6]. This approach offers an interpretable transition state (TS) model, such as the famous Houk–List model in proline catalysis[7,8], which allows the mechanistic rationalization of chirality control. Using the mechanistic knowledge of the stereoselectivity-determining TS model, one can engineer the chemical

[1]Center of Chemistry for Frontier Technologies, Department of Chemistry, State Key Laboratory of Clean Energy Utilization, Zhejiang University, Hangzhou, China. [2]Institut für Organische und Biomolekulare Chemie, Georg-August-Universität Göttingen, Göttingen, Germany. [3]Wöhler-Research Institute for Sustainable Chemistry (WISCh), Georg-August-Universität Göttingen, Göttingen, Germany. [4]Beijing National Laboratory for Molecular Sciences, Beijing, China. [5]Key Laboratory of Precise Synthesis of Functional Molecules of Zhejiang Province, School of Science, Westlake University, Hangzhou, China. ✉e-mail: Lutz.Ackermann@chemie.uni-goettingen.de; hxchem@zju.edu.cn

structure of the chiral catalyst with paper and pencil. However, the mechanism-based approach often faces challenges because the performance of asymmetric catalysis is simultaneously controlled by multiple elusive factors. A seemingly trivial structural variation in the substrate may lead to critical and non-intuitive change in the stereoselectivity outcome[9,10], which results in laborious and time-consuming screening efforts in current asymmetric catalysis research.

To improve the predictive capability towards asymmetric catalysis, a data-driven approach has emerged as a revolutionary strategy (Fig. 1a)[11–13]. This approach digitalizes the catalytic system in an informatics fashion and trains a statistical model to capture and predict the chemical pattern[14,15]. Landmark progress, which includes Sigman's multivariant linear regression studies[16,17] and Denmark's neural network modelling[18,19], reveals the remarkable potential of a data-driven approach in asymmetric catalysis[20,21]. Owing to the high-dimensional nature of the structure–stereoselectivity relationship, the statistical model is usually data hungry, and the ideal case of a complete dataset and big data support is rare in the catalysis field. In fact, the motivation and expense of catalysis research results in the available data being unfortunately limited and biased. This 'greedy' sampling leads to the incomprehensive knowledge of the synthetic space for known asymmetric transformations. Tedious reoptimizations of the catalytic conditions and modification of the catalyst structure are often necessary for customized applications[22], even if the asymmetric transformation is already optimized on selected model substrates. Therefore, the lack of a quantified and holistic understanding of the synthetic space hinders the practical applications of literature procedures, which limits the downstream availability of functional chiral compounds and materials.

Simply increasing the amount of training data may not solve the data-hunger problem of asymmetric catalysis prediction, given the unlimited dimensions of synthetic space[23–25] and the available efficiency to experiment or simulate an enantioselective transformation. It would be ideal to harness both the benefits of the mechanism-based approach (for example, interpretability and extrapolation ability) and the data-driven approach (for example, quantification and predictive ability). By implementing the knowledge of the TS in machine learning (ML), we surmise that the statistical model is able to take advantage of the chemist's interpretation of asymmetric catalysis and to translate it into digital language, to make the accurate and extrapolative prediction that can guide the synthetic application. In this work, we designed and developed a ML workflow that enables the holistic prediction of the synthetic space, which is demonstrated in palladium-catalysed electro-oxidative C–H bond activation reactions (Fig. 1b)[26]. The use of TS knowledge-based descriptors enabled an accurate ML prediction of enantioselectivity, whose extrapolation ability was verified by experimental tests. Analysis of the established ML model indicated the hidden contribution of the olefin to the enantioselectivity determination, and our density functional theory (DFT) calculations revealed the mechanistic origins of this olefin effect. Relying on the knowledge-based encoding and the design of the ensemble learning framework, the model was able to evaluate its reliability for each specific region of interest in the synthetic space and realized the quantitative examination of close to one million possibilities.

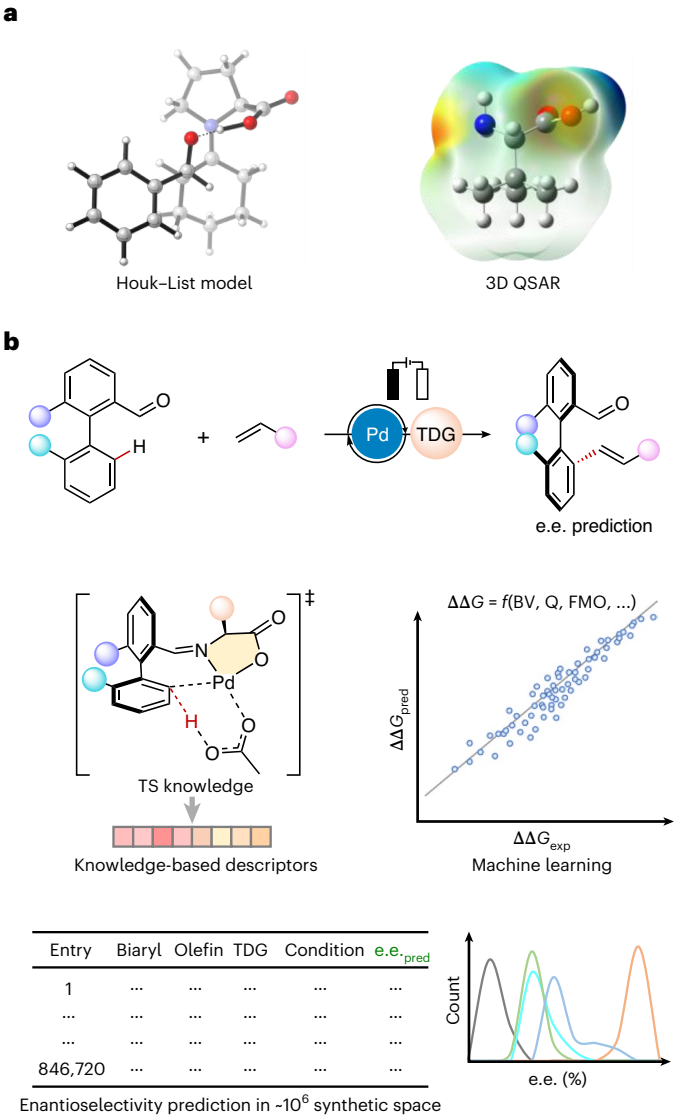


Fig. 1 | Prediction strategies for asymmetric catalysis. a, Previous strategies for asymmetric catalysis prediction, which include mechanism-based approaches (left) and data-driven approaches (right). **b**, Enantioselectivity prediction of pallada-electrocatalysed C–H activation (top) by implementing TS knowledge in ML (middle). Enabled by this ML model, a synthetic space of 846,720 possibilities was explored (bottom). $\Delta\Delta G = -RT\ln(\text{e.r.})$, where e.r. is the enantiomeric ratio, T is the reaction temperature and R is the gas constant. FMO, frontier molecular orbital descriptors; Q, atomic charge; QSAR, quantitative structure–activity relationship; exp, experiment; pred, prediction.

Results and discussion

Design of prediction workflow

To achieve the desired synthetic space prediction, we designed a workflow to harness the value of the reported catalysis data and TS knowledge (Fig. 2). The first step is to collect the reported catalysis results in a structured fashion. Owing to the greedy nature of catalysis development, the collected data are typically unbalanced in the label space of synthetic performance. Subsequent generation of the 'TS-like' distorted geometries provides the needed structural basis for the vectorization of TS knowledge. Step three is to generate a distinctive set of molecular descriptors using the TS-like geometries. A series of guidelines is proposed to implement the TS model in a descriptor design (vide infra), which improves the model's predictive ability towards the target transformation. The designed descriptors also provide a digital representation of the synthetic space and the distribution of the available samplings. Using this information, step four trains ML models to make enantioselectivity predictions and to evaluate the reliability of the predicted results. The base model aims to capture the general structure-performance relationship (SPR) by using the representative data in the explored synthetic space. For a specific region of the synthetic space, a delta model is also trained using the neighbouring data, so as to learn the regional perturbation of the general SPR. Through these ensemble predictions, one can explore the synthetic space of any possible combinations with a quantified synthetic performance and prediction reliability. These five steps can be practiced as a feedback loop if further experimentations are needed for the desired accuracy and confidence.

https://doi.org/10.1038/s44160-022-00233-y

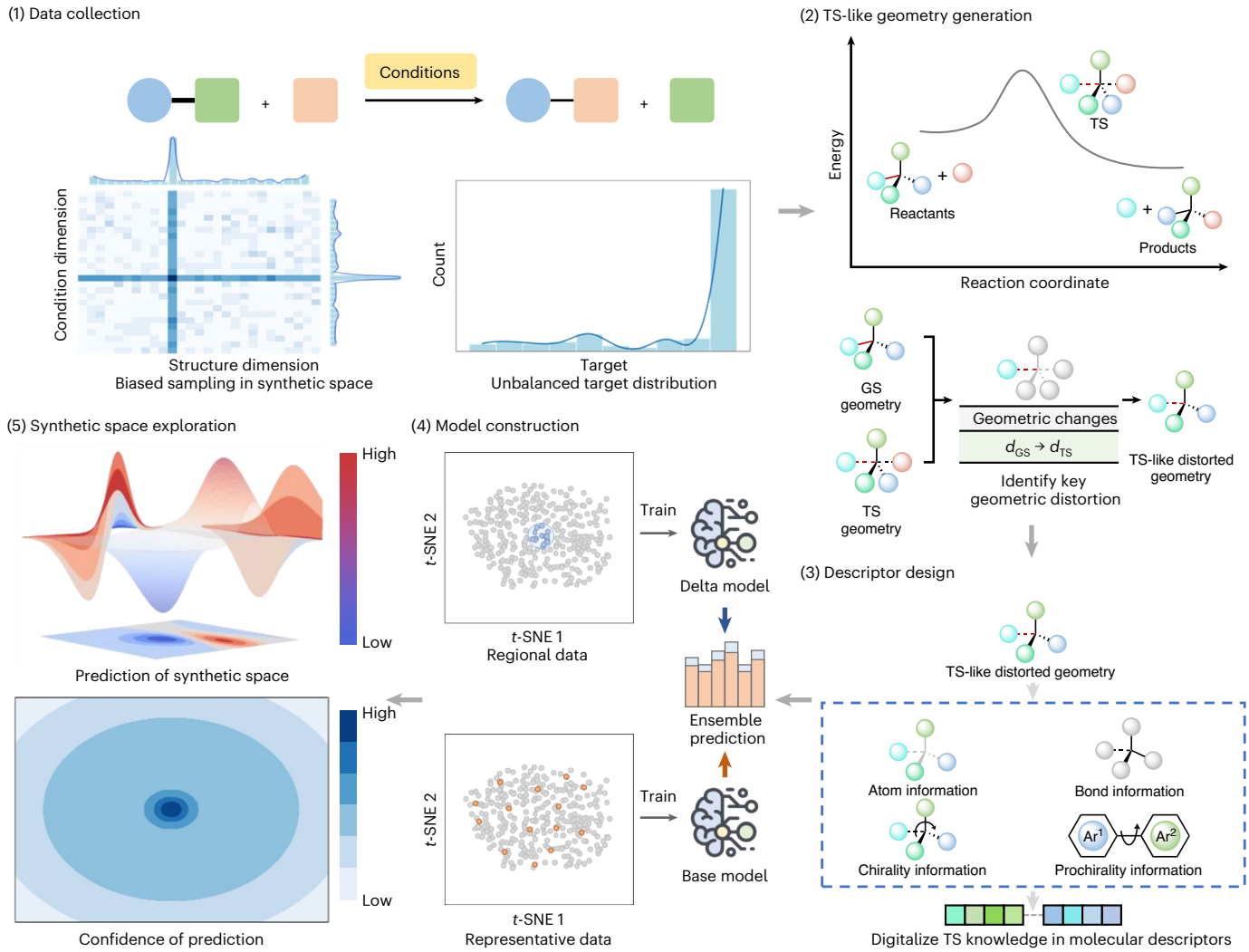


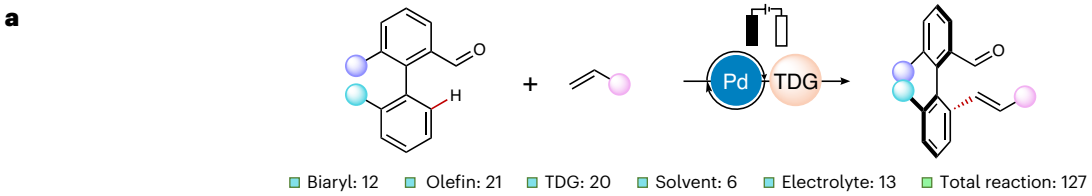
Fig. 2 | Workflow design for synthetic space prediction. Step (1): Collect the reaction data in a structured fashion. Owing to the greedy nature of synthetic exploration, the collected data are typically biased in synthetic space with the focused exploration of condition dimension (to identify the optimal condition) or structure dimension (to evaluate the substrate scope). This results in an unbalanced target distribution of the collected data, which does not fully represent the target distribution in reality. Step (2): This step generates the distorted geometry by a constrained optimization under the key distorted geometric parameters of the TS template. These distorted geometric parameters are identified by comparing ground state (GS) geometries with TS geometries. Step (3): Based on the TS-like geometries, design descriptors that vectorize the information of the reaction centre provide a distinctive set of molecular descriptors. Step (4): Train the base model using the representative data to capture the general SPR. Train the delta model using the regional data around the target transformation to capture the local perturbation of the SPR. The final ML prediction is the ensemble prediction using the base model and the delta model. Step (5): Using the ensemble prediction, all the possible transformations in the synthetic space can be evaluated with a quantified synthetic performance and prediction confidence. d_{GS}, bond length in the GS geometry; d_{TS}, bond length in the TS geometry.

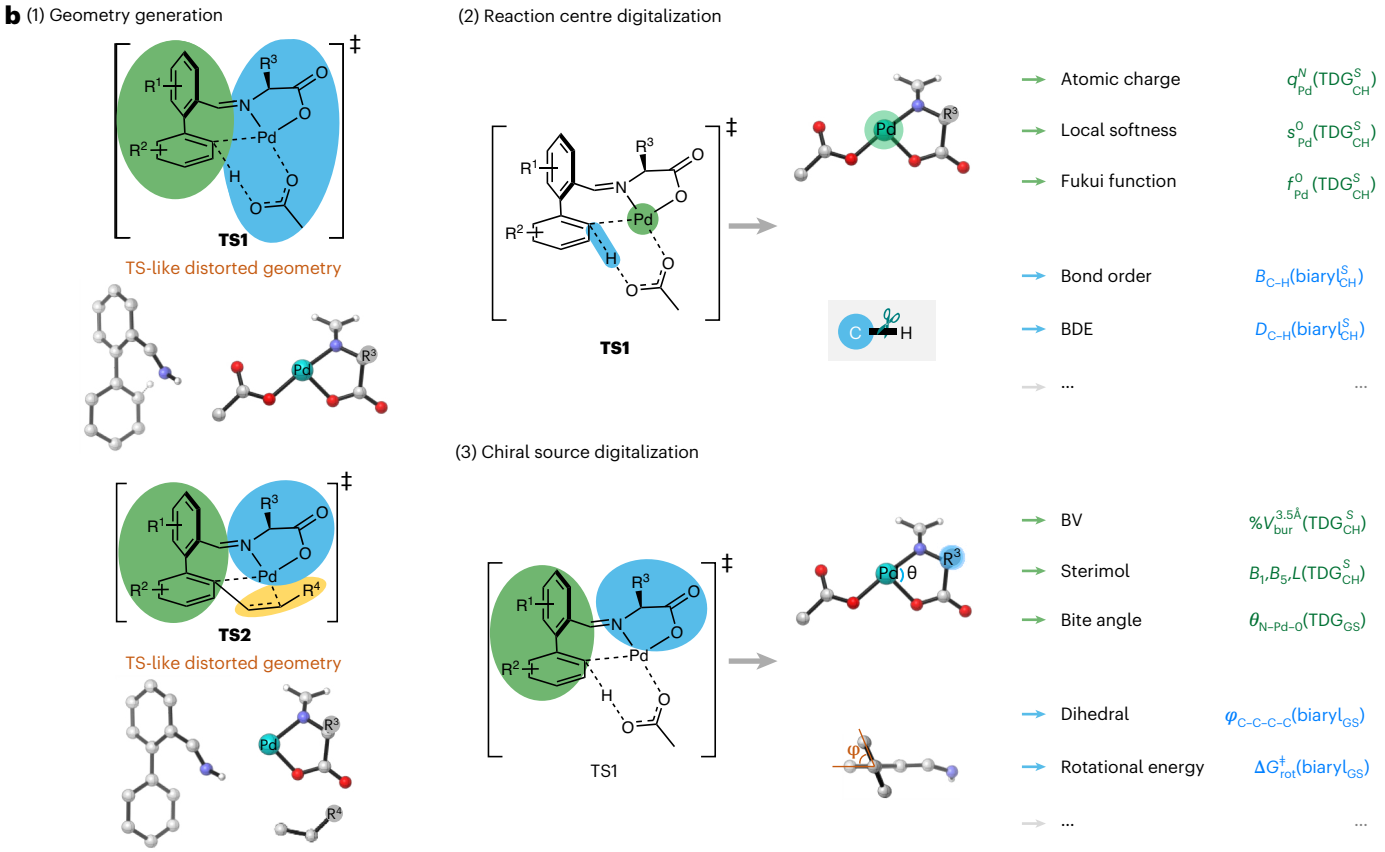
Development of the ML model

The essence of our ML modelling was to implement TS knowledge in the model training. Although the information of the TS was found useful in ML predictions of synthetic outcome by Grzybowski and co-workers[27] and Buttar[28] and co-workers, the implementation of TS knowledge in ML is still somewhat elusive. There is currently no universal approach to vectorize the understanding of a given TS model despite the extensive TS calculations in organic chemistry. From our experience of TS-based mechanistic studies[29], we propose a step-by-step guideline that can transfer a given TS model to key physical organic descriptors for reaction encoding, which is demonstrated in the base-assisted metallation-deprotonation TS, **TS1**, and the alkene insertion TS, **TS2**, in pallada-electrocatalysed C–H bond functionalization (Fig. 3).

The guideline includes three key components (Fig. 3b): (1) Generation of a TS-like distorted geometry. The TS model reflects a distorted geometric configuration that determines the reaction barrier, which provides a valuable encoding source for the ML model. We developed a subgraph recognition-based approach to automatically acquire the TS-like distorted geometries (Supplementary Fig. 5). For each TS, a universal geometric template was applied for all the transformations to avoid the combination explosion issue. Through this approach, **TS1** is represented by the distorted imine and the Pd–TDG (TDG, transient directing group) complex, and **TS2** is represented by a different set of the distorted imine and the Pd–TDG complex along with the distorted alkene. (2) Digitalization of the reaction centre. The TS model allows the identification of the cleaving and forming bonds in a chemical process, which usually involves the critical moiety for reactivity and selectivity determination. Coding the involved atoms and bonds offers useful site-specific features[30–32] to capture the influence from the reaction centre. (3) Digitalization of the chiral source. For stereoselectivity prediction, the chiral source is essential for the enantiodifferentiation and should

https://doi.org/10.1038/s44160-022-00233-y





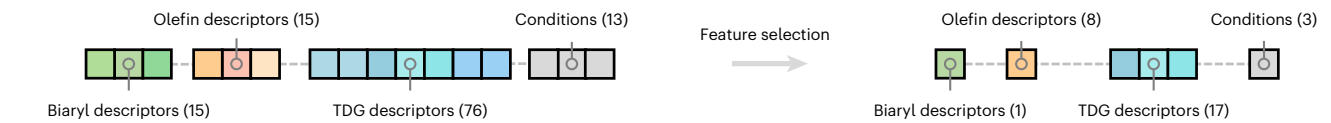


Fig. 3 | Dataset of the pallada-electrocatalysed C–H olefination and design of TS model-based encoding. a, Details of the collected transformation data. **b**, Designed guidelines to vectorize the TS models of C–H bond activation and alkene insertion. Step (1) provides the TS-like distorted geometries of aldehyde, olefin and TDG based on the C–H bond activation TS, **TS1**, and the alkene insertion TS, **TS2**. Step (2) is the reaction-centre digitalization step that generates the site-specific descriptors for atoms and bonds that are involved in the bond cleavage and formation. $q^{N}_{Pd}\left(TDG^{S}_{CH}\right)$, $s^{0}_{Pd}(TDG^{S}_{CH})$ and $f^{0}_{Pd}(TDG^{S}_{CH})$ are, respectively, the Hirshfeld charge, the condensed local softnesses and the condensed-to-atom Fukui function of palladium in the Pd–TDG complex derived from the (*S*)-C–H bond activation TS; $B_{C-H}(biaryl^{S}_{CH})$ and $D_{C-H}(biaryl^{S}_{CH})$ are, respectively, the Wiberg bond index and the homolytic bond dissociation energy (BDE) of the cleaving C–H bond of the biaryl aldehyde derived from the (*S*)-C–H bond-activation TS. Step (3) is the chiral source digitalization step that produces the descriptors that encode the chirality-related information of the Pd–TDG complex and biaryl aldehyde. $\%V^{3.5Å}_{bur}(TDG^{S}_{CH})$ is the BV parameter of the Pd–TDG complex derived from (*S*)-C–H bond activation TS; $B_1(TDG^{S}_{CH})$, $B_5(TDG^{S}_{CH})$ and $L(TDG^{S}_{CH})$ are the Sterimol parameters of the Pd–TDG complex derived from the (*S*)-C–H bond-activation TS; $\theta_{N-Pd-O}(TDG_{GS})$ is the bite angle of the TDG of the GS Pd–TDG complex; $\varphi_{C-C-C}(biaryl_{GS})$ and $\Delta G^{‡}_{rot}(biaryl_{GS})$ are the aryl–aryl dihedral angle and rotational free energy of the GS biaryl aldehyde, respectively. The complete details of the descriptors are provided in Supplementary Tables 2–4. **c**, Constitution of the reaction encoding prior to (left) and after (right) the feature selection.

be represented in the vectorization[12,33]. Using **TS1** as an example, the steric properties of the palladacycle species, as well as the prochiral elements, are recorded as descriptors. These three guidelines can, in principle, be applied to any TS model to provide a set of local descriptors, which digitalizes the domain knowledge of the target chemical process.

The collected data include 127 distinctive C–H olefination reactions that involve variations of 12 biaryls, 21 olefins, 20 TDGs, 6 solvents and 13 electrolytes. Details of the dataset are provided in Supplementary Figs. 1 and 2. Based on our previous mechanistic studies[26], the catalytic cycle of the palladium-catalysed electro-oxidative C–H bond activation involves a sequential base-assisted

https://doi.org/10.1038/s44160-022-00233-y

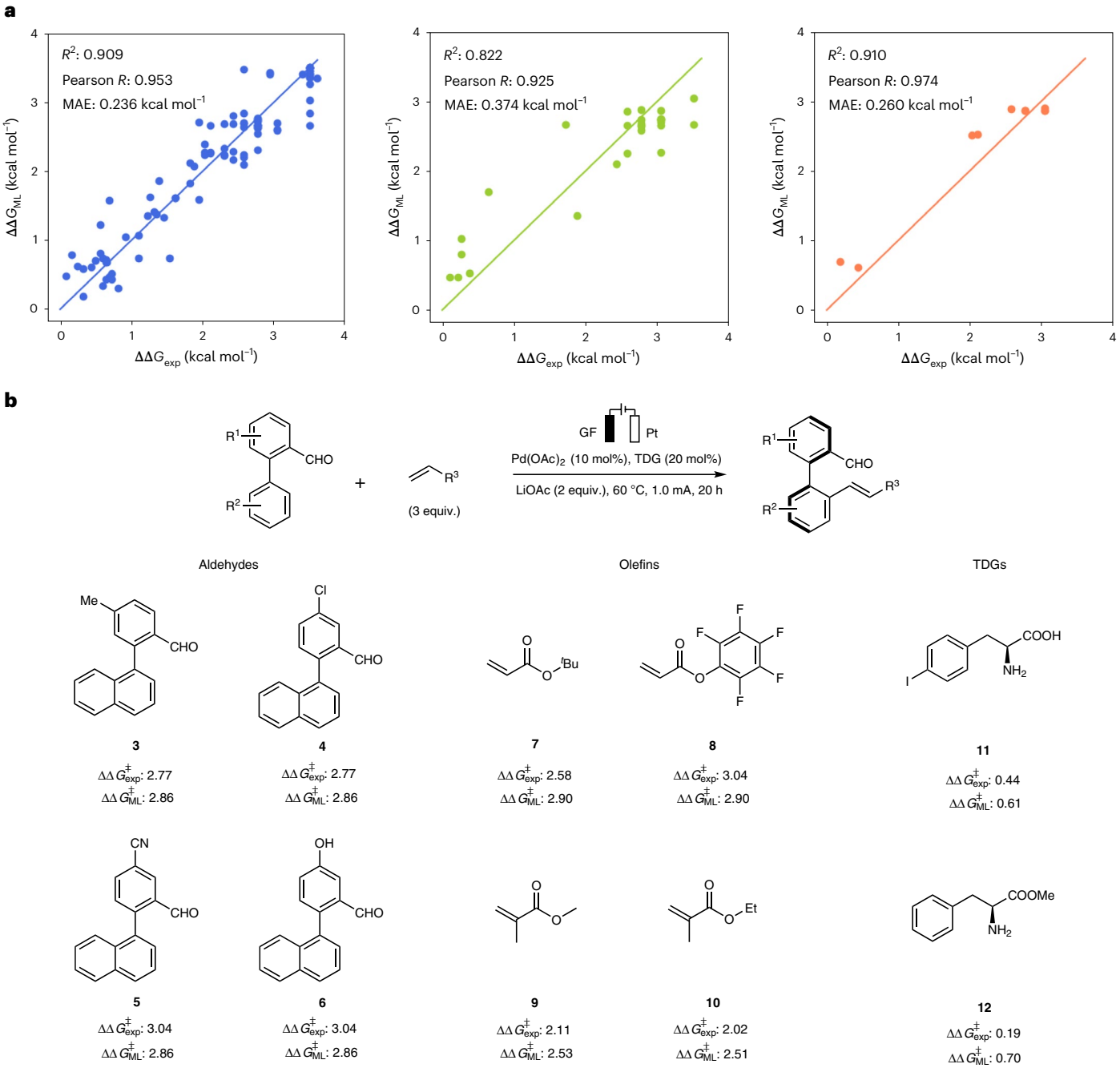


Fig. 4 | Regression performance of the designed ML model. a, Model performances in the tenfold cross-validation (left), OOS test set (centre) and external experimental tests (right). **b**, Details of the ML predictions and the verifications in the external experimental tests. GF, graphite felt.

metallation–deprotonation and alkene insertion TS, which irreversibly determines the product's axial chirality. Applying the above-designed guidelines to these two TS led to a 406-dimensional descriptor for the involved compounds. These descriptors, in addition to the encodings of the reaction conditions (solvent, electrolyte, temperature and current), provided a 419-digit vectorized representation of the palladaelectrocatalysed C–H bond activation (see Supplementary Fig. 7 and Supplementary Tables 1–5 for details of the generated descriptors). Subsequent correlation analysis removed the redundant molecular descriptors, and recursive feature elimination further deleted the non-essential encodings from 119 to only 29 dimensions while improving the model performance (Fig. 3c). Using these knowledge-based reaction encodings, ML training of the enantioselectivity regression identified ExtraTrees as the superior algorithm, whose results of the tenfold cross-validation are shown in Fig. 4a with a coefficient of determination (R^2) of 0.909 and a mean absolute error (MAE) of 0.236 kcal mol^{-1} (ML details are provided in Supplementary Figs. 9–11 and Supplementary Tables 6–8). For physical organic descriptors that have more than one generation method, such as the various forms of charge (for example, Hirshfeld, Mulliken and NPA (natural population analysis) charges) or bond order (for example, Wiberg and Mayer), we found that the change of generation approach has a limited influence on the model performance (Supplementary Table 12).

To further demonstrate the descriptive ability of the TS-based vectorization, we next evaluated the performances of the ML model in a series of extrapolation tests. In the collected dataset, 28 compounds

https://doi.org/10.1038/s44160-022-00233-y

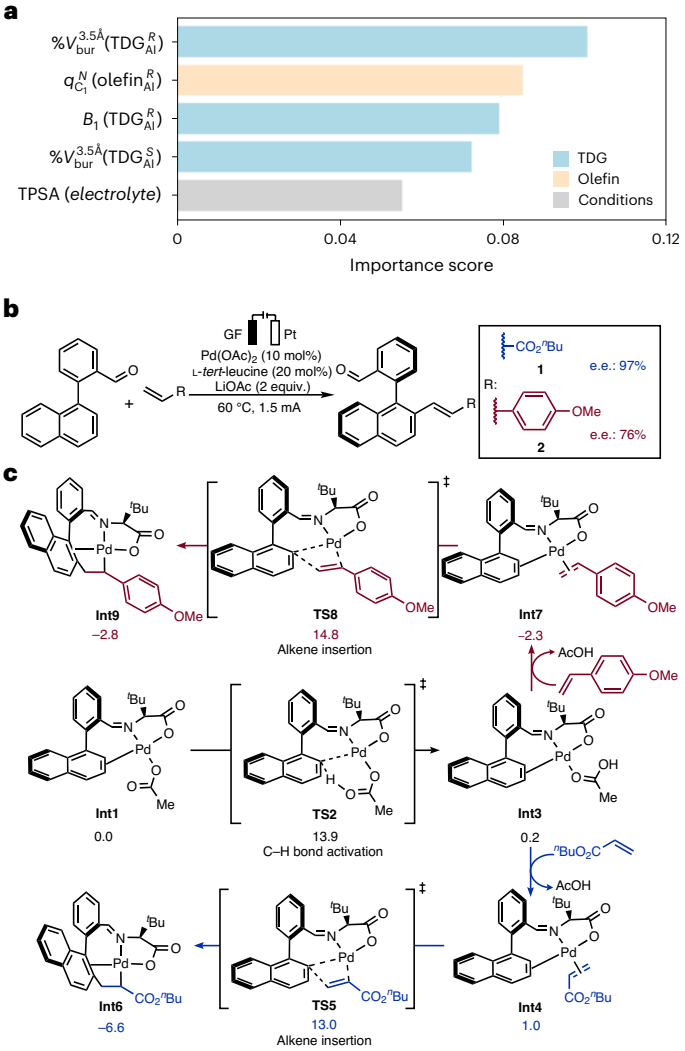


Fig. 5 | Model interpretation and mechanistic study of the olefin effect. **a**, Feature importance ranking: $\%V_{Bur}^{3.5Å}(TDG_{AI}^{R})$ and $\%V_{Bur}^{3.5Å}(TDG_{AI}^{S})$ are the BV parameters of the stereogenic carbon centre in the distorted Pd–TDG complex derived from the *R*- and *S*-alkene insertion TS, respectively; $q_{C1}^{N}(olefin_{AI}^{R})$ is the atomic charge of the terminal carbon of the distorted olefin derived from the *R*-alkene insertion TS and $B_1(TDG_{AI}^{R})$ is the Sterimol parameter B_1 of the distorted Pd–TDG complex derived from the *R*-alkene insertion TS. **b**, Experimental enantioselectivities of the pallada-electrocatalysed C–H olefination with *n*-butyl acrylate and 4-methoxystyrene. **c**, DFT-computed free-energy changes (kcal mol^{-1}) of the C–H bond activation and alkene insertion steps for *n*-butyl acrylate and 4-methoxystyrene. Int, intermediate.

(7 biaryls, 14 olefins and 7 TDGs) have only one associated transformation, and these transformations were selected out of the original dataset as the out-of-sample (OOS) test set. Although the OOS compounds were not seen during the model training, our model provided satisfying predictions with an R^2 of 0.822 and a MAE of 0.374 kcal mol^{-1} (Fig. 4a). These OOS predictions were not sensitive to the structure similarity between the training set compounds and the OOS compounds (Supplementary Fig. 13), which corroborates the extrapolation ability of the 'learned' structure–enantioselectivity relationship. Further comparisons confirmed that the TS encoding improves the model's predictive ability for the 'unseen' compounds in the OOS set; the widely applied molecular encodings, which include one-hot, RDKit descriptors, molecular fingerprints, many-body tensor representation and local many-body tensor representation, showed notably worse regression performances in the OOS test (R^2 ranged from 0.33 to 0.64; Supplementary Fig. 14). We believe that these high-dimensional molecular descriptors have challenges when handling a focused synthetic dataset with a limited size, which results in the unsatisfying extrapolation ability of the trained model. As an external test, we also synthesized a series of new biaryl aldehydes, olefins and TDGs to evaluate the model's predictive ability. The knowledge-based model performed well with a convincing correlation (R^2 of 0.910 and MAE of 0.260 kcal mol^{-1}; Fig. 4a). Comparing the ML-predicted and observed enantioselectivities (Fig. 4b), the model indeed captured the non-intuitive structural influence on the synthetic performance. It is particularly interesting that the model was able to identify the enantioselectivity mitigation in the 1,1-disubstituted alkenes **9** and **10**, which is against the original mechanistic model (vide infra) and challenging for human prediction.

Model interpretation and mechanistic study

Thanks to the chemical interpretability of the TS knowledge-based descriptors, ranking of the feature importance offers a data-driven perspective of the mechanistic origins of chirality control (Fig. 5a). Among the top-five descriptors, three of them belong to the steric properties ($\%V_{Bur}^{3.5Å}(TDG_{AI}^{R})$, $B_1(TDG_{AI}^{R})$ and $\%V_{Bur}^{3.5Å}(TDG_{AI}^{S})$; **AI**, alkene insertion). These buried volume (BV) and Sterimol parameters describe the steric environment of the Pd–TDG complex, which follows the mechanistic expectation that the amino acid is the chiral source of this transformation. In addition, the topological polar surface area (TPSA) of the electrolyte serves as the fifth important descriptor, and highlights the model's ability to capture the influence of electrocatalytic conditions on the enantioselectivity determination. It is particularly interesting that the ML model identified the charge of the terminal carbon of olefin ($q_{C1}^{N}(olefin_{AI}^{R})$) as the second important descriptor. This insight is against our previous mechanistic understandings because the DFT-computed free energy profile of the model substrates (2-(naphthalen-1-yl)benzaldehyde and *n*-butyl acrylate) suggested that the C–H bond activation is the enantioselectivity-determining step[26], and thus the olefin should not participate in the enantioselectivity determination.

In light of the discrepancy between the ML and DFT studies, we next explored the mechanistic origins of the olefin effect using DFT calculations. *n*-butyl acrylate **1** and 4-methoxystyrene **2** were selected as the model substrates due to the observed change of enantioselectivity (97 and 76% respectively; Fig. 5b). For *n*-butyl acrylate, the overall barrier of the C–H bond activation via **TS3** was 0.9 kcal mol^{-1} higher than that of the olefin insertion via **TS4**. This is consistent with our previous study[26] that the C–H bond activation is irreversible and the enantioselectivity-determining step, and thus the olefin substrate is not involved in the enantioselectivity determination. However, the DFT-computed free-energy profile of 4-methoxystyrene showed a reversed trend; this olefin has a lower insertion reactivity, and the overall barrier of the olefin insertion via **TS5** was 0.9 kcal mol^{-1} higher than that of the C–H bond activation via **TS3**. The detailed free-energy profiles are provided in Supplementary Fig. 37. These calculations revealed the mechanistic insight that the stereocontrol does not solely rely on the C–H activation step for all the viable olefinic substrates. For the case of the less effective olefins, the olefin insertion step requires a higher overall barrier compared with that of the C–H activation. This provides support for the olefin insertion step to be involved in the overall enantioselectivity determination of the catalytic transformation.

Based on the above results, it is noteworthy that neither the ML nor the DFT study is incorrect. In fact, the seeming discrepancy highlights the complementary features of the statistic and mechanistic approaches. ML is able to identify the hidden pattern in synthetic statistics, which could be challenging for the mechanistic study of representative systems. Further interpretation of the ML model and interpretation-driven mechanistic studies can improve the chemical understanding of the synthetic transformation, which provides the opportunity for a reinforcing feedback loop between knowledge and

https://doi.org/10.1038/s44160-022-00233-y

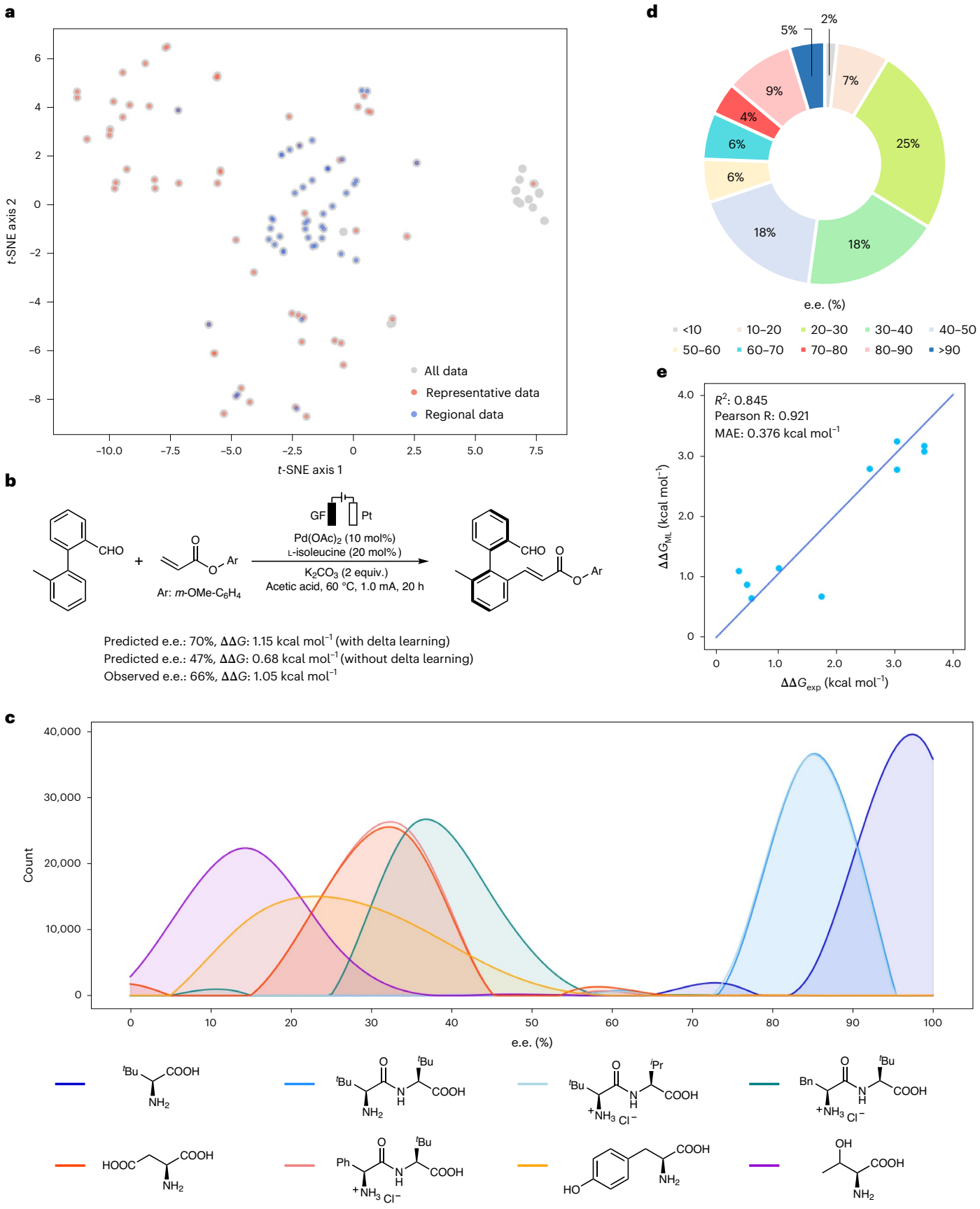


Fig. 6 | Synthetic space prediction and experimental verifications. **a**, *t*-SNE visualization of the experimentally sampled synthetic space and the transformation-specific data selection. **b**, Highlighted example of the prediction and comparison between ML predictions with or without delta learning. **c**, Distribution of predicted e.e. values for eight representative TDGs. **d**, Distribution in the studied synthetic space (top) of the predicted e.e. values (bottom) . **e**, Experimental verifications of selected synthetic space predictions.

data. Revealing the olefin effect on the enantioselectivity determination of the pallada-electrocatalysed C–H olefination highlighted the hidden mechanistic knowledge in the SPR data.

Prediction of synthetic space

The TS-model-based vectorization essentially projects the elusive synthetic space to a digitalized space composed of domain knowledge-related dimensions. Through this projection, one can comprehend the distribution of the performed experimental samplings in the entire synthetic space. Figure 6a is the *t*-distributed stochastic neighbour embedding (*t*-SNE) diagram of the sampled synthetic space, which maps the high-dimensional space to a two-dimensional (2D) figure to allow visualization. These experience-driven samplings were to achieve the highest catalysis performance and to evaluate the application scope, which occupies a highly biased fraction of the synthetic space. To achieve the holistic understanding of the synthetic space from this biased start, the evaluation of the model's reliability for each specific transformation in the space is necessary. This was realized by the design of a transformation-specific data selection and a delta learning strategy[34,35]. From the available statistics of the explored synthetic space, a representative dataset[36–38] was selected based on the cosine distance of the reaction encodings (red points; Fig. 6a). This representative dataset describes the general SPR of the explored space and supports the training of the base model. For a target prediction, the neighbouring regional data were selected to capture the local perturbation via delta learning and to evaluate the model's reliability.

An example of this prediction is presented in Fig. 6b. For this target prediction, the regional dataset of 45 'similar' transformations was selected (blue circles; Fig. 6a). These transformations predicted enantioselectivities from the base model and from the observed enantioselectivities; the differences in these values allowed the training of the delta learning model, which aimed to correct the base model's predictions and improve the model's reliability towards the target prediction. Using this delta learning strategy, the ensemble model (base prediction + delta prediction) gave a predicted $\Delta\Delta G^{\ddagger}$ of 1.15 kcal mol^{-1}, whereas the naive model training (using all available data without delta learning) gave a prediction of 0.68 kcal mol^{-1}. Our experimental verification of this transformation is 1.05 kcal mol^{-1}, which highlights the improving effect of the delta learning strategy. In this case, the beneficial delta learning suggests that the available statistics include the needed data for the local perturbation, and further experimental samplings around this transformation are not necessary. For cases that lack regional data, and thus the delta learning is not effective, the iterative ML-driven experimental samplings and model trainings would allow an automatic synthetic space exploration without human intervention, which suits the application of robotic synthesis.

Using the designed ensemble learning strategy, the entire synthetic space of 846,720 possibilities was explored. This massive space includes 5,040 combinations from the variations of biaryls, olefins and TDGs, and the additional 168 possibilities that result from a change of solvent, electrolyte and current. The ML predictions presented a quantified and holistic comprehension of the synthetic space, which is challenging for experience-driven experimental samplings. The distributions of the ML-predicted values for eight representative TDGs are shown in Fig. 6c. This gives direct knowledge of the statistical performance of each amino acid. Based on the predicted e.e. values (Fig. 6d), it should be noted that the high enantioselectivity examples only occupy a very small portion of the synthetic space. Solely 5% of the transformations have a predicted e.e. value higher than 90%, which again emphasizes the biased experimental samplings and data distribution in catalysis development. We also carefully examined these predictions to exclude the possibility of a simple statistical majority guess and confirmed the model's out-of-range predictive ability (Supplementary Table 16). To further verify the model's reliability and accuracy in the synthetic space prediction, ten random cases, which cover the predicted range of e.e. values, were tested experimentally. The ML predictions agreed well with the experimental observations (Fig. 6e); the model achieved an R^2 of 0.845 and a MAE of 0.376 kcal mol^{-1}, an accuracy that can provide solid support for the target-orientated exploration of the candidate transformations in the synthetic space.

Conclusion

In summary, we designed a workflow to achieve a ML prediction of synthetic performance using data from catalysis development. This workflow takes advantage of the domain knowledge of the TS model and the predictive ability of the ML approach, which together create a general strategy to make accurate and reliable synthetic performance predictions using data from biased experimental samplings. This ML strategy, which can be readily implemented in daily synthetic development scenarios, offers a bridge from the widely existing screening data and TS models in synthetic transformations to the desired holistic knowledge of a massive synthetic space. As demonstrated in palladium-catalysed electro-oxidative C–H bond activation reactions, the TS models of the C–H bond activation and the alkene insertion were vectorized, which supported the establishment of a ML model that achieved excellent enantioselectivity predictions with chemical accuracy. The model's reliability was further corroborated by a series of OOS tests and additional experimental verifications. Model interpretations disclosed the critical role of the olefin on the enantioselectivity prediction. This led to more in-depth mechanistic studies, which revealed that the rate-determining step, indeed, can vary depending on the olefin substitution pattern. For low-reactivity olefins, the olefin insertion step can be involved in the overall enantioselectivity determination, with the energy barrier for olefin insertion being higher than that of the C–H bond activation. These interpretation-driven mechanistic insights highlight the synergy between knowledge-based ML and a conventional reaction mechanism study.

This model realized the efficient exploration of the synthetic space with 846,720 possibilities. A transformation-specific data selection approach was designed to evaluate the model's reliability in the prediction of each transformation. These predictions presented a quantified and holistic comprehension of the synthetic space, which is extremely challenging for experience-driven experimental samplings. The accuracy of the prediction results was further examined by additional experimental tests, which featured an excellent agreement with the predicted results. This study provides a universal data-driven approach to harness the hidden value in catalysis screening data and TS knowledge, which will allow target-orientated optimizations in digitalized synthetic spaces. Note that the current approach applied the same geometric template for each TS model to avoid a combination explosion issue. Thus, the change of TS position is not captured. Further advancement of the automated TS location technique[39] would enable the generation of substrate-specific TS geometries, which can enrich the structural sources of our approach. For transformations without clear TS structures, we believe that the same ML approach can be effective through the ensemble learning of elementary transformation models. Based on the TS structures of known elementary transformations, ML can in principle learn the most determining elementary transformation from synthetic statistics. This will enable designed TS-based ML without the requirement for a mechanistic model of the target transformation. Related works are currently under investigation in our laboratories.

Methods

Details of ML

Generation of TS-like distorted geometry. A subgraph recognition-based approach was developed to acquire the TS-like distorted geometries for the involved reaction components in an automatic and efficient fashion (Supplementary Fig. 5). Based on the TS geometry and the GS geometry for a given compound, the customized script first detected the shared fragment of the two 3D coordinate files using a predefined

subgraph. Subsequent comparisons of all the geometric parameters in the shared fragment quantified the geometric changes from the GS geometry to the corresponding fragment in the TS. Applying a customized threshold ($|\Delta d| > 0.1$ Å, $|\Delta\theta| > 15°$ or $|\Delta\varphi| > 30°$) identified the key geometric parameters that differentiate the ground state and TS, which resulted in the template geometric parameters. Subsequent constrained optimizations by imposing these geometric parameters provided the TS-like geometries. Further details are provided in Supplementary Figs. 5 and 6).

Descriptor generation. For a TS-like distorted geometry, a series of site-specific features were calculated for the atoms and bonds involved in the bond cleavage and formation process (Supplementary Fig. 7). This applied to the TS-like distorted biaryls, olefins and TDGs, which digitalized the information of the reaction centre. In addition to the reaction centre, the information of the chiral source of TDGs and biaryls was encoded to support the ML prediction of enantioselectivity (Supplementary Fig. 7). To encode the reaction conditions, physical organic features were used to describe the solvents and electrolytes. The full list of generated physical organic descriptors is provided in Supplementary Tables 1–5. Additionally, a few widely applied molecular descriptors were also generated and tested in the ML modelling, which included one hot encoding, 2D descriptors and 3D descriptors. The generation details of these descriptors are provided in Supplementary Table 7.

Model training. For the training of the ML models, hyperparameter optimization was performed to identify the optimal hyperparameter settings. A tenfold cross-validation was used to test the performance of the combinations of candidate ML algorithms and molecular descriptors. By comparing the MAE, Pearson R and R^2 in the tenfold cross-validation, the regression performances were evaluated and identified the suitable ML algorithm for the TS-based descriptors. Subsequently, feature selection was performed to decrease the complexity of the ML model using recursive feature elimination with cross-validation. Details of the tested ML algorithms are provided in Supplementary Table 6.

Synthetic space prediction. A modelling workflow of ensemble prediction was designed for the synthetic space exploration. The base model was trained by the representative data of the sampled synthetic space. For each target transformation in the synthetic space, a delta model was trained using the available regional data around this transformation. During the delta model training, the regional dataset had the true label from the experimental observation and the predicted label from the base model. The delta value between these two labels allowed the training of the delta model. The final enantioselectivity prediction of the synthetic space is the sum of the predicted values from the base model and the delta model. The full details of the ensemble prediction procedure of synthetic space are provided in Supplementary Fig. 15.

Details of experiment

In the general procedure for atroposelective pallada-electrocatalysed C–H olefination, the electrocatalysis was carried out in an undivided cell, equipped with a graphite felt anode and a Pt cathode. Biaryls (0.20 mmol, 1.0 equiv.), acrylates (3 equiv.), $Pd(OAc)_2$ (10 mol%), TDG (20 mol%) and additive (2 equiv.) were placed in an undivided cell and dissolved in 4.5 ml of solvent. Electrocatalysis was performed at 60 °C with a constant current of 1.0 mA for 20 h. At ambient temperature, the reaction mixture was diluted with EtOAc. After removal of the solvent in vacuo, the crude mixture was purified by column chromatography on silica gel to yield the products.

Data availability

Data related to ML details, experimental procedures, HPLC spectra and NMR spectra are available in the Supplementary Information. Source data are provided with this paper.

Code availability

Codes for target transformation, descriptor generation, model training, feature selection, feature ranking and synthetic space exploration are freely available at https://github.com/licheng-xu-echo/SyntheticSpacePrediction.

References

1. Noyori, R. Asymmetric catalysis: science and opportunities (Nobel Lecture). *Angew. Chem. Int. Ed.* **41**, 2008–2022 (2002).
2. Trost, B. M. Asymmetric catalysis: an enabling science. *Proc. Natl Acad. Sci. USA* **101**, 5348–5355 (2004).
3. Noyori, R. Synthesizing our future. *Nat. Chem.* **1**, 5–6 (2009).
4. Taylor, M. S. & Jacobsen, E. N. Asymmetric catalysis in complex target synthesis. *Proc. Natl Acad. Sci. USA* **101**, 5368–5373 (2004).
5. Woodard, S. S., Finn, M. G. & Sharpless, K. B. Mechanism of asymmetric epoxidation. 1. Kinetics. *J. Am. Chem. Soc.* **113**, 106–113 (1991).
6. Cheong, P. H.-Y., Legault, C. Y., Um, J. M., Çelebi-Ölçüm, N. & Houk, K. N. Quantum mechanical investigations of organocatalysis: mechanisms, reactivities, and selectivities. *Chem. Rev.* **111**, 5042–5137 (2011).
7. Bahmanyar, S., Houk, K. N., Martin, H. J. & List, B. Quantum mechanical predictions of the stereoselectivities of proline-catalyzed asymmetric intermolecular aldol reactions. *J. Am. Chem. Soc.* **125**, 2475–2479 (2003).
8. Lam, Y.-h, Grayson, M. N., Holland, M. C., Simon, A. & Houk, K. N. Theory and modeling of asymmetric catalytic reactions. *Acc. Chem. Res.* **49**, 750–762 (2016).
9. Knowles, R. R. & Jacobsen, E. N. Attractive noncovalent interactions in asymmetric catalysis: links between enzymes and small molecule catalysts. *Proc. Natl Acad. Sci. USA* **107**, 20678–20685 (2010).
10. Neel, A. J., Milo, A., Sigman, M. S. & Toste, F. D. Enantiodivergent fluorination of allylic alcohols: data set design reveals structural interplay between achiral directing group and chiral Anion. *J. Am. Chem. Soc.* **138**, 3863–3875 (2016).
11. Crawford, J. M., Kingston, C., Toste, F. D. & Sigman, M. S. Data science meets physical organic chemistry. *Acc. Chem. Res.* **54**, 3136–3148 (2021).
12. 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).
13. Oliveira, J. C. A. et al. When machine learning meets molecular synthesis. *Trends Chem.* **4**, 863–885 (2022).
14. Mater, A. C. & Coote, M. L. Deep learning in chemistry. *J. Chem. Inf. Model.* **59**, 2545–2559 (2019).
15. Tkatchenko, A. Machine learning for chemical discovery. *Nat. Commun.* **11**, 4125 (2020).
16. Niemeyer, Z. L., Milo, A., Hickey, D. P. & Sigman, M. S. Parameterization of phosphine ligands reveals mechanistic pathways and predicts reaction outcomes. *Nat. Chem.* **8**, 610–617 (2016).
17. Reid, J. P. & Sigman, M. S. Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature* **571**, 343–348 (2019).
18. Zahrt, A. F. et al. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* **363**, eaau5631 (2019).
19. Henle, J. J. et al. Development of a computer-guided workflow for catalyst optimization. Descriptor validation, subset selection, and training set analysis. *J. Am. Chem. Soc.* **142**, 11578–11592 (2020).
20. Singh, S. et al. 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).

21. Gallarati, S. et al. Reaction-based machine learning representations for predicting the enantioselectivity of organocatalysts. *Chem. Sci.* **12**, 6879–6889 (2021).
22. Kutchukian, P. S. et al. Chemistry informer libraries: a chemoinformatics enabled approach to evaluate and advance synthetic methods. *Chem. Sci.* **7**, 2604–2613 (2016).
23. Hase, F., Roch, L. M., Kreisbeck, C. & Aspuru-Guzik, A. Phoenics: a Bayesian optimizer for chemistry. *ACS Cent. Sci.* **4**, 1134–1145 (2018).
24. Coley, C. W. Defining and exploring chemical spaces. *Trends Chem.* **3**, 133–145 (2021).
25. Shields, B. J. et al. Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **590**, 89–96 (2021).
26. Dhawa, U. et al. Enantioselective pallada-electrocatalyzed C–H activation by transient directing groups: expedient access to helicenes. *Angew. Chem. Int. Ed.* **59**, 13451–13457 (2020).
27. Moskal, M., Beker, W., Szymkuc, S. & Grzybowski, B. A. Scaffold-directed face selectivity machine-learned from vectors of non-covalent interactions. *Angew. Chem. Int. Ed.* **60**, 15230–15235 (2021).
28. Jorner, K., Brinck, T., Norrby, P.-O. & Buttar, D. Machine learning meets mechanistic modelling for accurate prediction of experimental activation energies. *Chem. Sci.* **12**, 1163–1175 (2021).
29. Zhang, S. Q. & Hong, X. Mechanism and selectivity control in Ni- and Pd-catalyzed cross-couplings involving carbon–oxygen bond activation. *Acc. Chem. Res.* **54**, 2158–2171 (2021).
30. Tomberg, A., Johansson, M. J. & Norrby, P. O. A predictive tool for electrophilic aromatic substitutions using machine learning. *J. Org. Chem.* **84**, 4695–4703 (2019).
31. Guan, Y. et al. Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors. *Chem. Sci.* **12**, 2198–2208 (2020).
32. Li, X., Zhang, S. Q., Xu, L. C. & Hong, X. Predicting regioselectivity in radical C–H functionalization of heterocycles through machine learning. *Angew. Chem. Int. Ed.* **59**, 13253–13259 (2020).
33. Gallegos, L. C., Luchini, G., St John, P. C., Kim, S. & Paton, R. S. Importance of engineered and learned molecular representations in predicting organic reactivity, selectivity, and chemical properties. *Acc. Chem. Res.* **54**, 827–836 (2021).
34. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Big data meets quantum chemistry approximations: the delta-machine learning approach. *J. Chem. Theory Comput.* **11**, 2087–2096 (2015).
35. Xu, L. C. et al. Towards data-driven design of asymmetric hydrogenation of olefins: database and hierarchical learning. *Angew. Chem. Int. Ed.* **60**, 22804–22811 (2021).
36. Martin, T. M. et al. Does rational selection of training and test sets improve the outcome of QSAR modeling? *J. Chem. Inf. Model.* **52**, 2570–2578 (2012).
37. 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).
38. Rinehart, N. I., Zahrt, A. F., Henle, J. J. & Denmark, S. E. Dreams, false starts, dead ends, and redemption: a chronicle of the evolution of a chemoinformatic workflow for the optimization of enantioselective catalysts. *Acc. Chem. Res.* **54**, 2041–2054 (2021).
39. Dewyer, A. L., Argüelles, A. J. & Zimmerman, P. M. Methods for exploring reaction space in molecular systems. *WIREs Comput. Mol. Sci.* **8**, e1354 (2018).

Acknowledgements

Generous support by the National Natural Science Foundation of China (21873081 and 22122109, X. Hong; 22103070, S.-Q.Z.), the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-SIAS-006, X. Hong), Beijing National Laboratory for Molecular Sciences (BNLMS202102, X. Hong), CAS Youth Interdisciplinary Team (JCTD-2021-11, X. Hong), Fundamental Research Funds for the Central Universities (226-2022-00140 and 226-2022-00224, X. Hong), the Center of Chemistry for Frontier Technologies and Key Laboratory of Precise Synthesis of Functional Molecules of Zhejiang Province (PSFM 2021-01, X. Hong), the State Key Laboratory of Clean Energy Utilization (ZJUCEU2020007, X. Hong), China Scholarship Council (fellowship to X. Hou), the European Union (ERC advanced grant no. 101021358 conferred to L.A.) and the DFG (Gottfried-Wilhelm-Leibniz-Preis attributed to L.A. and SPP 2363) are gratefully acknowledged. Calculations and ML trainings were performed on the high-performance computing system at the Department of Chemistry, Zhejiang University.

Author contributions

X. Hong and L.A. conceived and supervised the project. X. Hong and S.-Q.Z. designed the workflow of the ML. L.-C.X. and S.-W.L. performed the ML training and analysed the training data. J.F. and X. Hou performed the experiments and analysed the experimental data. Y.-Y.L. and J.C.A.O. performed the DFT calculations for the physical organic descriptors and the mechanistic studies. X. Hong. and L.A. wrote the manuscript with input from all the authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44160-022-00233-y.

Correspondence and requests for materials should be addressed to Lutz Ackermann or Xin Hong.

Peer review information *Nature Synthesis* thanks Tobias Gensch, Bartosz Grzybowski and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Peter Seavill, in collaboration with the *Nature Synthesis* team.