Instytut Chemii Organicznej
Polskiej Akademii Nauk

ROZPRAWA DOKTORSKA

Metody uczenia maszynowego w doborze katalizatorów oraz przewidywaniu reaktywności i enancjoselektywności w katalizie asymetrycznej kompleksami magnezu

mgr Paulina Baczewska

Monotematyczny cykl publikacji wraz z komentarzem przedstawiony Radzie Naukowej Instytutu Chemii Organicznej Polskiej Akademii Nauk w celu uzyskania stopnia doktora

Promotor: Prof. Jacek Młynarski

Warszawa, 2026

W tym miejscu pragnę podziękować:

Prof. Jackowi Młynarskiemu

Za bycie mentorem i nieocenioną pomoc w realizacji badań, a przede wszystkim za poczucie, że przez cały czas graliśmy w tej samej drużynie. Za wsparcie i gotowość do pomocy w sytuacjach, w których sama nie wiedziałam, co zrobić. Za owocne dyskusje, współpracę w pozytywnej atmosferze oraz motywujące słowa w chwilach zwątpienia.

Zespołowi XIII – byłym i obecnym członkom:

Joannie Jaszczewskiej-Adamczak – za ogromną ilość pozytywnej energii, którą ładowała laboratorium, świetną atmosferę i za to, że zawsze wiedziała gdzie jest każdy odczynnik w Instytucie. I oczywiście za owocne współprace i niezawodność

Łukaszowi Włoszczakowi, Tomkowi Podlewskiemu, Robertowi Pawłowskiemu Oskarowi Popikowi i Marcie Turkiewicz – za niekończące się dyskusje i prawdziwie dekadenckie dywagowanie, które były jedyne w swoim rodzaju i na pewno poszerzyły moją perspektywę

Marinie Bellido – za godziny rozmów, które działały terapeutycznie oraz dzielenie się podnoszącą na duchu filozofią życia

Kolegom i koleżankom z zaprzyjaźnionych zespołów w IChO – za świetne wspólne spotkania na terenie Instytutu, długie rozmowy, koleżeńskość i wsparcie, dzięki którym IChO jest miejscem, które ciepło będę wspominać

Dziękuję wszystkim, z którymi realizowałam projekty badawcze – za zaangażowanie i pomoc

Moim najbliższym przyjaciołom

Idze i Michałowi – mojej podporze, na którą zawsze mogłam liczyć

Siostrze Patrycji – która niezmiennie zawsze stoi po mojej stronie i wspiera

Robertowi i Szymonowi – za bycie w moim życiu przez te wszystkie lata

Kubie – za kolory, które wprowadził do mojego życia podczas pisania tej dysertacji

Dziękuję mojej rodzinie za słuchanie, cierpliwość i wiarę we mnie.

FINANSOWANIE

Badania finansowane w ramach grantu:



NARODOWE CENTRUM NAUKI

OPUS NCN 2021/41/B/ST4/01617

„Zastosowanie sztucznej inteligencji do katalizy asymetrycznej: nowa droga do projektowania katalizatorów”

Spis treści

Spis publikacji wchodzących w skład rozprawy doktorskiej 3
Spis pozostałych publikacji 5
Spis wystąpień konferencyjnych 6
Streszczenie w języku polskim 7
Streszczenie w języku angielskim / Abstract in English 8
Wykaz skrótów i pojęć 9
PRZEWODNIK PO ROZPRAWIE DOKTORSKIEJ 13
Założenia i cel pracy 13
Cele szczegółowe 13
Układ rozprawy 14
Wstęp literaturowy 15
1. Wstęp 15
1.1 Chiralność, synteza asymetryczna 15
1.2 Strategie otrzymywania związków enancjomerycznie wzbogaconych 15
1.3 Kataliza asymetryczna i skala energetyczna problemu 17
1.4 Jak projektuje się katalizatory 18
1.5 Dlaczego wyznaczanie enancjoselektywności drogą obliczeniową jest wyzwaniem? 19
2. Kataliza asymetryczna z wykorzystaniem kompleksów magnezu 20
2.1 Właściwości kompleksów magnezu istotne dla katalizy 20
2.2 Zakres transformacji i klasy chiralnych ligandów 21
2.3 Magnez jako atrakcyjny metal w katalizie i syntezie 21
3. Jak buduje się i ocenia model predykcyjny – kompendium do zrozumienia rozprawy 22
3.1 Od zapisu reakcji do wektora liczbowego 22
3.2 Co należy modelować: ΔΔG‡ czy %*ee* 25
3.3 Rozkład danych i zakres stosowalności 25
3.4 Podział danych i walidacja 26
3.5 Metryki oceny modelu 28
3.6 Interpretowalność modeli i analiza istotności cech 28
4. Uczenie maszynowe w katalizie asymetrycznej 29
4.1 Zakres rozdziału 29
4.2 Korelacja — czy selektywność można zapisać jako funkcję struktury? 30
4.3 Transferowalność — czy wiedzę z jednej reakcji można wykorzystać w innej? 36
4.4 Projektowanie katalizatorów — czy model może wskazać nową strukturę katalizatora? 42
4.5 Wybór katalizatora z wcześniej zdefiniowanej przestrzeni 46

4.6 Generalizacja — czy model uczy się zależności chemicznych, czy tylko odtwarza znane dane? 52
5. Podsumowanie stanu wiedzy 57
5.1 Stan wiedzy w momencie rozpoczęcia badań 57
5.2 Problem badawczy I. Czy rozproszona literatura może służyć do rekomendowania katalizatora? 57
5.3 Problem badawczy II. Jakie pytania można stawiać małym zbiorom laboratoryjnym? 58
5.4 Wspólne ograniczenie obu podejść. Generalizacja i walidacja eksperymentalna 58
5.5 Późniejsza literatura jako potwierdzenie kierunku 59
Szczegółowy kontekst projektów i badania własne 60
6.1 Rekomendacja układu katalitycznego na podstawie danych literaturowych 60
6.1.1 Budowa zbioru danych 60
6.1.2 Reprezentacja numeryczna reakcji i warunków 63
6.1.3 Model, jego optymalizacja i ocena 64
6.1.4 Analiza istotności cech oraz wpływ warunków reakcji 66
6.1.5 Walidacja eksperymentalna 67
6.1.6 Podsumowanie 73
6.2 Przewidywanie reaktywności i enancjoselektywności na podstawie małych zbiorów danych eksperymentalnych na przykładzie reakcji katalizowanych kompleksami magnezu 75
6.2.1 Układ Mg–ProPhenol i reakcje modelowe 75
6.2.2 Budowa zbiorów danych 76
6.2.3 Rozdzielenie zadania predykcyjnego 77
6.2.4 Reprezentacja chemiczna i dobór modeli 78
6.2.5 Wybór zmiennej odpowiedzi 81
6.2.6 Walidacja eksperymentalna 81
6.2.7 Zakres stosowalności i granice predykcji 82
6.2.8 Addycja tio-Michaela 84
6.2.9 Podsumowanie 85
7. Podsumowanie i wnioski 87
Bibliografia 89
PUBLIKACJE ORYGINALNE I OŚWIADCZENIA AUTORÓW 97
8. Machine Learning Algorithm Guides Catalyst Choices for Magnesium-Catalyzed Asymmetric Reactions 97
9. Predicting Reactivity and Enantioselectivity from Small Experimental Datasets: A Case Study in Asymmetric Magnesium Catalysis 107
10. Oświadczenia Autorów 121

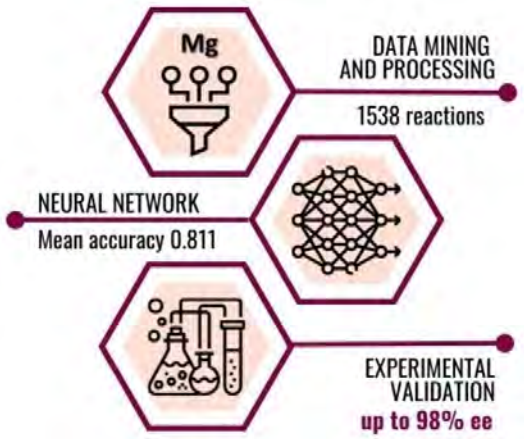
Spis publikacji wchodzących w skład rozprawy doktorskiej

Podstawą niniejszej rozprawy są dwie publikacje oryginalne:

1 | P. Baczewska, M. Kulczykowski, B. Zambroń, J. A. Jaszczewska-Adamczak, Z. Pakulski, R. Roszak, B. A. Grzybowski, J. Mlynarski, Machine Learning Algorithm Guides Catalyst Choices for Magnesium-Catalyzed Asymmetric Reactions, *Angew. Chem. Int. Ed*. **2024**, 63, e202318487.

DOI: 10.1002/anie.202318487 (https://doi.org/10.1002/anie.202318487)

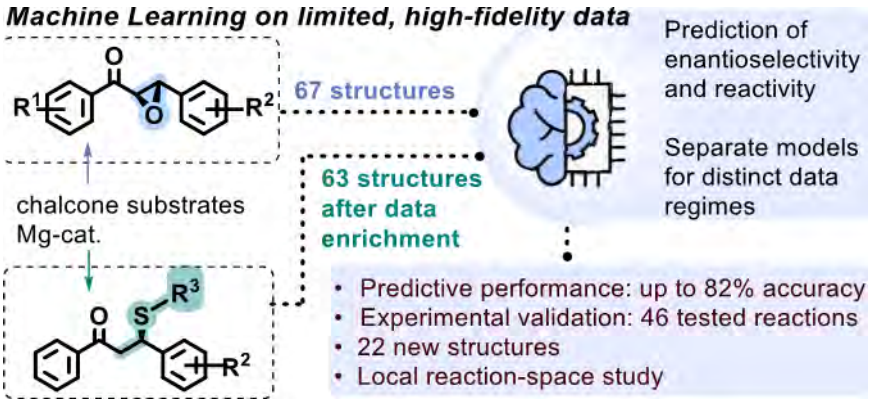
Abstrakt graficzny:



2 | Paulina Baczewska, Damian Nowak, Joanna Jaszczewska-Adamczak, Rafał A. Bachorz, Marcin Hoffmann, Jacek Mlynarski; Predicting Reactivity and Enantioselectivity from Small Experimental Datasets: A Case Study in Asymmetric Magnesium Catalysis. *ACS Catal.* **2026**.

DOI: 10.1021/acscatal.6c05090 (https://doi.org/10.1021/acscatal.6c05090)

Abstrakt graficzny:



Spis pozostałych publikacji

Publikacje oryginalne:

3 | J. A. Jaszczewska-Adamczak, P. Baczewska, R. Bujok, J. Mlynarski, Magnesium-Catalyzed Asymmetric Thia-Michael Addition to α,β-Unsaturated Ketones, *Adv. Synth. Catal.* **2024**, 366, 1412–1421. DOI: 10.1002/adsc.202301414

4 | M. Michalak, M. Nowacki, P. Baczewska, S. Chakraborty, W. Depa, Z. Rafiński, K. Śniady-Maciążek, O. Danylyuk, P. Cmoch, M. Zimnicka, K. Niemirowicz-Laskowska, B. Skonieczna, H. Car, W. Kośnik, A. Saha, A. Poater, Chiral NHC–Au(III) complexes: catalysis, biology, and tosylate-controlled enantioselectivity, *Inorg. Chem. Front*. **2026**. DOI: 10.1039/d6qi01280a

Praca przeglądowa:

5 | P. Baczewska, K. Śniady, W. Kośnik, M. Michalak, Acenaphthene-Based N-Heterocyclic Carbene Metal Complexes: Synthesis and Application in Catalysis, *Catalysts* **2021**, 11, 972. DOI: 10.3390/catal11080972

Spis wystąpień konferencyjnych

Wyniki przedstawione w niniejszej rozprawie doktorskiej zaprezentowano na następujących konferencjach:

1 | Niemiecko-Polska Konferencja Chemii Organicznej w Sopocie, 20–24.09.2026 – referat ustny: Machine Learning-Guided Catalyst Selection in Asymmetric Magnesium and Zinc Catalysis

2 | Związki Biologicznie Czynne: aktywność, struktura, synteza, Białystok, 26–28.06.2024 — referat ustny: Zastosowanie modelu uczenia maszynowego do optymalizacji katalizatorów w asymetrycznych reakcjach katalizowanych magnezem

3 | Zjazd Wiosenny Sekcji Młodych Polskiego Towarzystwa Chemicznego, Poronin, 1–5.05.2024 — prezentacja posterowa: Analiza kontroli asymetrycznej w reakcji tio-Michaela katalizowanej dinuklearnym kompleksem magnezu i (*S*,*S*)-ProPhenolu

Streszczenie w języku polskim

Tempo, w jakim przybywa danych chemicznych, dawno przekroczyło możliwości ich ręcznego przeglądania. Wiedza zawarta w opublikowanych reakcjach jest rozproszona po setkach artykułów i materiałów uzupełniających, a zależności widoczne dopiero na poziomie większego zbioru mogą łatwo pozostać niezauważone.

Właśnie w tym miejscu widzę rolę uczenia maszynowego. Nie jako wartości samej w sobie, lecz jako narzędzia pozwalającego wydobyć z istniejących wyników zależności użyteczne dla katalizy asymetrycznej, która nadal pozostaje obszarem pełnym wyzwań.

Niniejsza rozprawa doktorska opiera się na założeniu, że o wartości modelu decyduje to, czy pytanie zadane modelowi odpowiada informacji rzeczywiście zawartej w danych i ich reprezentacji. Nie można na przykład pytać o wpływ typu rozpuszczalnika na wynik reakcji, jeżeli w zbiorze nie ma przykładów tej samej reakcji prowadzonej w różnych rozpuszczalnikach. Pytania badawcze stawiałam w obszarze asymetrycznej katalizy kompleksami magnezu.

Pierwszy projekt pokazuje pracę z dużym i niejednorodnym zbiorem danych, którego samo ujednolicenie było istotnym wyzwaniem. Rozproszoną literaturę katalizy magnezowej przekształciłam w spójny zbiór danych, obejmujący również wyniki o niskim nadmiarze enancjomerycznym oraz przykłady niekorzystne, odzyskane ręcznie z materiałów uzupełniających. Zbudowany na tym zbiorze model wskazywał ligandy, które w połączeniu z odpowiednią solą magnezową miały największą szansę dać indukcję asymetryczną dla zadanej reakcji. Jego wartość potwierdziłam eksperymentalnie. Rekomendowane układy poprawiły wyniki znanych redukcji α-ketoestrów, pozwoliły otrzymać produkty, dla których wcześniej nie opisano redukcji katalizowanej kompleksem magnezu, a w części reakcji umożliwiły zastąpienie labilnego odczynnika magnezoorganicznego wygodniejszą w użyciu i tańszą solą magnezu.

Drugi projekt dotyczył sytuacji odwrotnej. Punktem wyjścia były małe, ale jednorodne zbiory danych dla dwóch reakcji opracowanych wcześniej w naszym laboratorium, czyli epoksydacji chalkonów i addycji tio-Michaela. Dane zostały uzupełnione o rzadko publikowane wyniki negatywne. W takim przypadku pytanie ilościowe bywa możliwe, ale nie zawsze. W epoksydacji chalkonów wartości enancjoselektywności były do siebie zbliżone, a brak produktu nie mógł być traktowany jako 0% ee, dlatego sens miało przede wszystkim pytanie o reaktywność. W addycji tio-Michaela, obejmującej szerszą przestrzeń substratów i większe zróżnicowanie wyników, możliwe było przewidywanie nadmiaru enancjomerycznego oraz uchwycenie nieoczywistych trendów. Oba typy predykcji sprawdziłam prospektywnie na substratach niewykorzystanych do budowy modeli.

Obie prace prowadzą do wspólnego wniosku. O użyteczności modelu nie decyduje wyłącznie liczba danych ani złożoność algorytmu, lecz również to, czy postawione pytanie mieści się w granicach informacji zawartej w zbiorze. Tak rozumiany model nie zastępuje eksperymentu. Pomaga wskazać, co warto sprawdzić w pierwszej kolejności i gdzie predykcji należy ufać ostrożnie. Dlatego każdą rekomendację traktowałam jako hipotezę wymagającą weryfikacji doświadczalnej. Uczenie maszynowe może w ten sposób przekształcić opublikowaną już wiedzę w dobrze uzasadnione punkty wyjścia do optymalizacji metod katalizy asymetrycznej.

Streszczenie w języku angielskim / Abstract in English

The pace at which chemical data are being generated has long exceeded the limits of manual analysis. The knowledge contained in published reactions is scattered across hundreds of articles and supporting information files, and relationships that become visible only at the level of a larger dataset can easily remain unnoticed.

This is where I see the role of machine learning. Not as a value in itself, but as a tool that makes it possible to extract from existing results relationships that are useful for asymmetric catalysis, which remains a field full of challenges.

This dissertation is based on the assumption that the value of a model depends on whether the question asked of it matches the information actually contained in the data and in their representation. For example, one cannot ask about the effect of solvent on the reaction outcome if the dataset does not contain examples of the same reaction performed in different solvents. I formulated these research questions in the area of asymmetric catalysis with magnesium complexes.

The first project shows how one can work with a large and heterogeneous dataset, for which the process of standardization itself was a major challenge. I transformed the scattered literature on magnesium catalysis into a consistent dataset that also included low-enantiomeric-excess results and unfavorable examples manually recovered from supporting information. The model built on this dataset indicated ligands that, in combination with an appropriate magnesium salt, had the highest chance of inducing enantioselectivity in a given reaction. I confirmed its value experimentally. The recommended catalytic systems improved the outcomes of known α-ketoester reductions, enabled access to products for which no magnesium-complex-catalyzed reduction had previously been reported, and in some reactions made it possible to replace a labile organomagnesium reagent with a more convenient and less expensive magnesium salt.

The second project concerned the opposite situation. The starting point was a set of small but internally consistent datasets for two reactions previously developed in our laboratory, namely chalcone epoxidation and thia-Michael addition. These datasets were supplemented with rarely published negative results. In this setting, we studied when a quantitative question can be asked. In chalcone epoxidation, the enantioselectivity values were close to one another, and the absence of product could not be treated as 0% ee, so the most meaningful question concerned reactivity. In the thia-Michael addition, which covered a broader substrate space and a wider range of outcomes, it was possible to predict enantiomeric excess and capture non-obvious trends. I tested both types of predictions prospectively on substrates that had not been used to build the models.

Both studies lead to the same conclusion. The usefulness of a model is determined not by the number of data points alone, nor by the complexity of the algorithm, but by whether the question being asked falls within the limits of the information contained in the dataset. Understood in this way, the model does not replace experiment. It helps indicate what should be tested first and where predictions should be treated with caution. For this reason, I treated every recommendation as a hypothesis requiring experimental verification. In this way, machine learning can turn accumulated literature knowledge into informed starting points for catalyst discovery.

Wykaz skrótów i pojęć

Uczenie maszynowe, statystyka i ocena modeli:

AUPRC — pole pod krzywą precyzja–czułość (ang. area under the precision–recall curve)

F1 — miara F1 — średnia harmoniczna precyzji i czułości

LORO — walidacja z pominięciem jednej reakcji (ang. leave-one-reaction-out)

MAE — średni błąd bezwzględny (ang. mean absolute error)

MLR — regresja wieloliniowa (ang. multiple linear regression)

PLS — regresja częściowych najmniejszych kwadratów (ang. partial least squares)

Protonet — sieć prototypowa — metoda meta-uczenia (ang. prototypical network)

QSAR — ilościowa zależność struktura–aktywność (ang. quantitative structure–activity relationship)

QSPR — ilościowa zależność struktura–właściwość (ang. quantitative structure–property relationship)

QSRR — ilościowa zależność struktura–retencja (ang. quantitative structure–retention relationship)

QSSR — ilościowa zależność struktura–selektywność (ang. quantitative structure–selectivity relationship)

RMSE — pierwiastek błędu średniokwadratowego (ang. root mean square error)

R^2 — współczynnik determinacji

SHAP — ang. SHapley Additive exPlanations — metoda oceny istotności cech

SVR — regresja wektorów nośnych (ang. support vector regression)

UMAP — metoda redukcji wymiarowości (ang. uniform manifold approximation and projection)

Reprezentacje i deskryptory molekularne:

ASO — średnie zatłoczenie steryczne (ang. average steric occupancy)

CGR — skondensowany graf reakcji (ang. condensed graph of reaction)

ChemBERTa — model językowy generujący osadzenia struktur (nazwa własna)

ChyLine — deskryptory fragmentowe ChyLine (nazwa własna)

CircuS — deskryptory podstruktur cyklicznych (ang. circular substructures)

CoMFA — analiza porównawcza pól molekularnych (ang. comparative molecular field analysis)

DRFP — różnicowy fingerprint reakcji (ang. differential reaction fingerprint)

ECFP — rozszerzony fingerprint łączności (ang. extended-connectivity fingerprint)

MBTR — ang. many-body tensor representation

SLATM — ang. spectrum of London and Axilrod–Teller–Muto representation

SMART — ang. Spatial Molding for Approachable Rigid Targets

SMILES — liniowy zapis struktury (ang. simplified molecular-input line-entry system)

Sterimol — parametry steryczne Sterimol (nazwa własna)

TSRC — stan przejściowy wychwytu rodnika (ang. transition state of radical capture)

TSRE — stan przejściowy eliminacji redukcyjnej (ang. transition state of reductive elimination)

UTS — uniwersalny zbiór treningowy (ang. universal training set)

Pojęcia i techniki chemiczne:

ΔΔG‡ — różnica energii swobodnej aktywacji

DFT — teoria funkcjonału gęstości (ang. density functional theory)

LUMO — najniższy nieobsadzony orbital molekularny (ang. lowest unoccupied molecular orbital)

NHC — karben N-heterocykliczny (ang. N-heterocyclic carbene)

NMR — magnetyczny rezonans jądrowy (ang. nuclear magnetic resonance)

SAD — asymetryczne dihydroksylowanie Sharplessa (ang. Sharpless asymmetric dihydroxylation)

Ligandy, katalizatory i reagenty:

AD-mix — mieszanina do asymetrycznego dihydroksylowania (ang. asymmetric dihydroxylation mixture)

BINOL — 1,1′-bi-2-naftol

BOX — ligand bisoksazolinowy

COD — 1,5-cyklooktadien (ang. 1,5-cyclooctadiene)

CPA — chiralny kwas fosforowy (ang. chiral phosphoric acid)

HBpin — pinakoloboran (4,4,5,5-tetrametylo-1,3,2-dioksaborolan)

IDPi — katalizator imidodifosforanowy (ang. imidodiphosphorimidate)

n-Bu_2Mg — di-n-butylomagnez

NADH — zredukowany dinukleotyd nikotynoamidoadeninowy

NTf_2 — bis(trifluorometanosulfonylo)imid

OTf — triflan (grupa trifluorometanosulfonianowa)

ProPhenol — ligand ProPhenol (nazwa własna)

TCYP — chiralny kwas fosforowy stosowany jako katalizator (nazwa własna)

THF — tetrahydrofuran

TRIP — chiralny kwas fosforowy 3,3′-Bis(2,4,6-triisopropylphenyl)-1,1′-binaphthyl-2,2′-diyl hydrogenphosphate

Narzędzia i bazy danych:

ChEMBL — baza danych związków bioaktywnych (nazwa własna)

MolVS — biblioteka do normalizacji struktur (nazwa własna)

RDKit — biblioteka cheminformatyczna (nazwa własna)

Reaxys — baza danych reakcji chemicznych (nazwa własna)

PRZEWODNIK PO ROZPRAWIE DOKTORSKIEJ

Założenia i cel pracy

Współczesna literatura chemiczna dostarcza stale rosnącej liczby publikacji, z których każda może zawierać wiele nowych wyników reakcyjnych. Wraz ze wzrostem liczby dostępnych danych coraz trudniejsze staje się ich manualne łączenie, rozpoznawanie zależności pomiędzy wynikami pochodzącymi z różnych prac oraz wyprowadzanie na tej podstawie bardziej ogólnych wniosków. W tej sytuacji metody oparte na danych, w tym uczenie maszynowe, stanowią naturalne uzupełnienie tradycyjnej analizy literatury, umożliwiając systematyczne poszukiwanie zależności w dużych zbiorach wyników eksperymentalnych. Przystępując do realizacji badań przyjęłam założenie, że informacje zgromadzone w literaturze mogą służyć nie tylko jako źródło pojedynczych precedensów syntetycznych, lecz również jako dane pozwalające poszerzać wiedzę o reakcjach i wspierać ich optymalizację. Podstawowym wymogiem jest to, aby dane odzwierciedlały informacje niezbędne do udzielenia odpowiedzi na pytanie badawcze oraz aby pytanie to mieściło się w przestrzeni chemicznej objętej tymi danymi. Podejście takie może być szczególnie użyteczne w obszarach o znacznym potencjale syntetycznym, które nadal wymagają szerszej eksploracji. Za taki obszar uważam katalizę asymetryczną z wykorzystaniem kompleksów magnezu, która stała się przedmiotem badań w moim doktoracie.

Jako tezę mojej pracy doktorskiej przyjęłam, że rodzaj i struktura dostępnych danych determinują, jakie pytanie predykcyjne można wiarygodnie postawić modelowi uczenia maszynowego, oraz że, przy świadomym uwzględnieniu ograniczeń wynikających z zakresu i jakości tych danych, modele takie mogą wspierać rozwój nowych metod katalitycznych oraz retrospektywną analizę zależności obecnych w zgromadzonych wynikach eksperymentalnych.

Cele szczegółowe

Celem pracy było zbadanie zakresu, w jakim metody uczenia maszynowego mogą wspierać decyzje podejmowane podczas rozwoju asymetrycznych reakcji katalizowanych kompleksami magnezu. W ramach pracy:

(1) zebrałam i opracowałam bazę reakcji katalizowanych kompleksami magnezu, obejmującą również reakcje odzyskane ręcznie z materiałów uzupełniających prac źródłowych, w tym przykłady o niskiej selektywności nieobecne w repozytoriach reakcji, oraz uczestniczyłam w budowie modelu rekomendującego chiralny ligand i warunki dla zadanej reakcji, odpowiadając za jego stronę chemiczną, ocenę poprawności merytorycznej kolejnych rozwiązań oraz proponowanie zmian w sposobie konstrukcji i interpretacji modelu (rozdział 6.1);
(2) zweryfikowałam wskazania tego modelu doświadczalnie zarówno dla reakcji wcześniej opisanych w asymetrycznej katalizie magnezowej, sprawdzając możliwość opracowania alternatywnego, skutecznego układu katalitycznego, jak i dla przemian dotąd niestosowanych w katalizie kompleksami magnezu, oceniając możliwość rozszerzenia zakresu tej katalizy na nowe typy reakcji. Walidacja obejmowała asymetryczną redukcję α-ketoestrów, w tym syntezę prekursora związku BMS-270394, redukcję pochodnych izatyny oraz addycję Michaela (rozdział 6.1);
(3) określiłam, o co można wiarygodnie pytać model zbudowany na małym, wewnętrznie spójnym zbiorze pochodzącym z jednego laboratorium, wykorzystując dwie

komplementarne transformacje katalizowane kompleksami magnezu (epoksydację chalkonów i addycję tio-Michaela). Odpowiadałam za przygotowanie i merytoryczną weryfikację danych oraz współtworzyłam założenia procesu modelowania, w tym rozdzielenie zadania klasyfikacji reaktywności od regresji enancjoselektywności. Uzyskane predykcje interpretowałam z uwzględnieniem zakresu stosowalności modeli i weryfikowałam prospektywnie dla wcześniej niebadanych substratów (rozdział 6.2).

Układ rozprawy

Rozdział 1. wprowadza problem doboru katalizatora w katalizie asymetrycznej. Rozdział 2. przedstawia chemię kompleksów magnezu w zakresie niezbędnym do zrozumienia obu prac. Rozdział 3. wprowadza pojęcia i miary uczenia maszynowego, ograniczone do tych, które w rozprawie faktycznie występują. Rozdział 4. przedstawia przegląd zastosowań uczenia maszynowego w katalizie asymetrycznej, uporządkowany według rosnącej wagi stawianych pytań. Rozdział 5. wyprowadza z tego przeglądu brakujące obszary wiedzy, oddzielając stan wiedzy sprzed rozpoczęcia badań od literatury późniejszej. Rozdział 6. opisuje badania własne spisane w publikacjach stanowiących podstawę niniejszej rozprawy doktorskiej: podrozdział 6.1 dotyczy modelu zbudowanego na danych literaturowych, a 6.2 modeli zbudowanych na małych zbiorach laboratoryjnych.

Wstęp literaturowy

1. Wstęp

1.1 Chiralność, synteza asymetryczna

Chiralność oraz możliwość kontrolowanego otrzymywania związków wzbogaconych optycznie należą do fundamentalnych zagadnień współczesnej chemii organicznej. Zastosowanie metod stereoselektywnej syntezy wykorzystywane jest bezpośrednio w chemii medycznej, agrochemii oraz projektowaniu materiałów funkcjonalnych.[1, 2, 3]

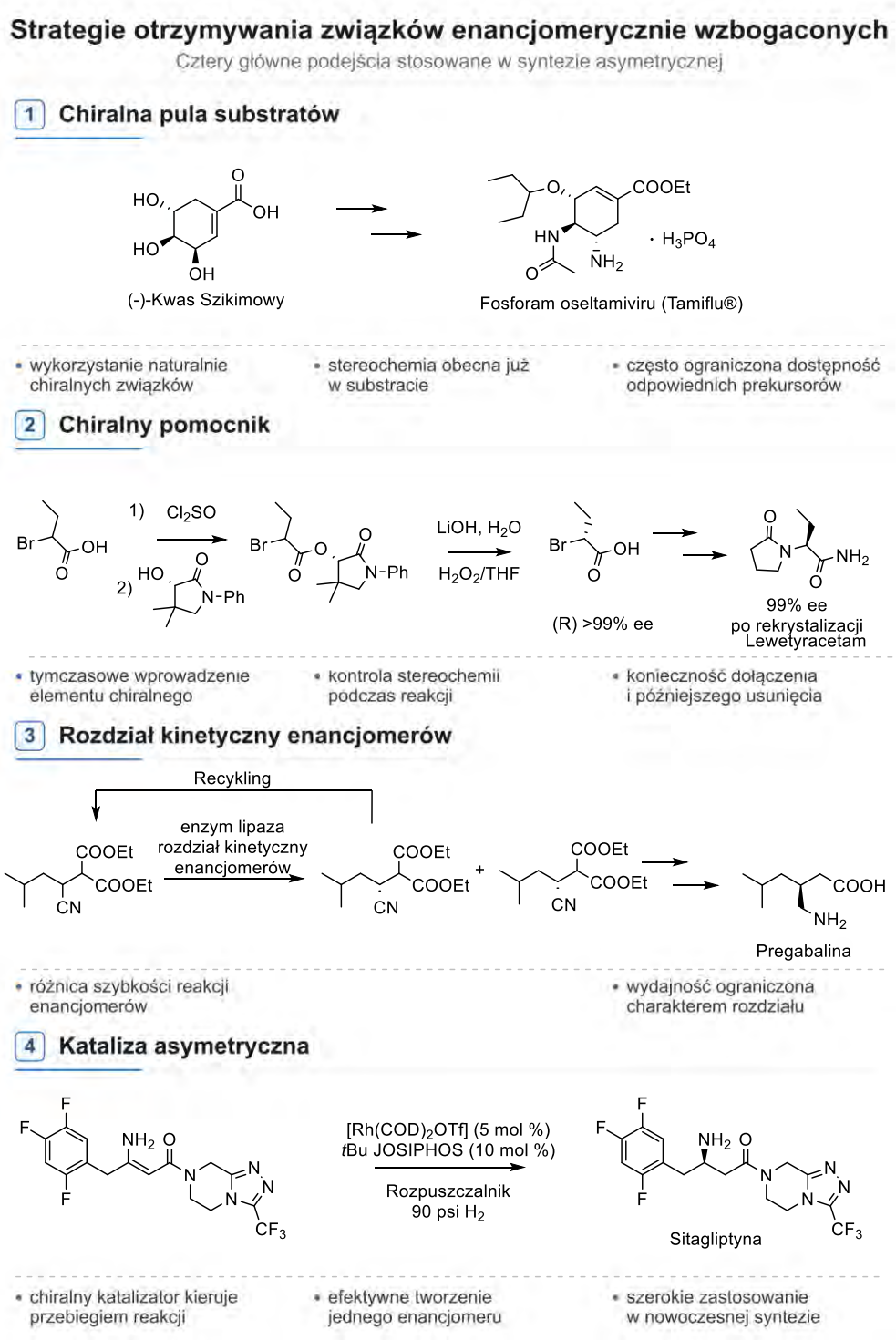
Obecnie ponad połowa stosowanych substancji leczniczych to związki chiralne, a nowo rejestrowane leki coraz częściej występują w postaci pojedynczych enancjomerów zamiast mieszanin racemicznych.[4] Tendencja ta wynika zarówno z odmiennych właściwości biologicznych poszczególnych enancjomerów, jak i z zaostrzenia wymagań regulacyjnych dotyczących bezpieczeństwa oraz aktywności farmakologicznej.[5] W konsekwencji rozwój wydajnych metod otrzymywania związków enancjomerycznie czystych pozostaje jednym z najważniejszych kierunków rozwoju syntezy organicznej.

Spośród nowych małocząsteczkowych substancji czynnych dopuszczonych do obrotu przez amerykańską Agencję Żywności i Leków w latach 2013–2022 pojedyncze enancjomery stanowiły 163 z 278, a więc 59%.[5] Efektywne metody syntezy asymetrycznej przekładają się zatem bezpośrednio na dostępność i koszt wytwarzania substancji leczniczych.

1.2 Strategie otrzymywania związków enancjomerycznie wzbogaconych

Związki optycznie czynne otrzymuje się czterema zasadniczymi drogami (Rysunek 1). Jedną z najbardziej opłacalnych ekonomicznie jest wykorzystanie związków naturalnych, które występują w postaci czystej enancjomerycznie, czyli korzystanie z chiralnej puli substratów (*chiral pool*). Przykładem strategii *chiral pool* jest synteza oseltamiwiru z naturalnego (−)-kwasu szikimowego, którego enancjomerycznie czysty szkielet cykloheksenowy stanowi źródło informacji chiralnej przenoszonej do produktu końcowego.[6] Inna metoda polega na wprowadzeniu do cząsteczki grupy zawierającej centra stereogeniczne, będącej tak zwanym pomocnikiem chiralnym, oraz przeprowadzeniu reakcji różnicowania stereochemicznego, w wyniku której na powstałe centrum stereogeniczne grupa chiralna dodana do cząsteczki ma kierujący wpływ (indukuje przez mechanizm reakcji i powstanie odpowiedniego centrum w konfiguracji *syn* lub *anti*). Jako przykład zastosowania pomocnika chiralnego można podać diastereoselektywną α-alkilację pochodnej imidu Evansa w syntezie lewetyracetamu. Po utworzeniu centrum stereogenicznego pomocnik jest usuwany z produktu.[7] Nieco mniej ekonomiczną metodą (w zależności od naszych celów) jest rozdział mieszaniny racemicznej – klasyczny lub kinetyczny. Przykładem zastosowania enzymatycznego rozdziału kinetycznego w procesie wytwarzania leku jest chemoenzymatyczna synteza pregabaliny opracowana przez Martíneza i współpracowników. W kluczowym etapie racemiczny 2-karboksyetylo-3-cyjano-5-metyloheksanian etylu poddaje się selektywnej hydrolizie z użyciem komercyjnej lipazy (Lipolase z *Thermomyces lanuginosus*). Jeden z enancjomerów ulega szybszej hydrolizie do kwasu karboksylowego, podczas gdy drugi pozostaje w postaci enancjomerycznie wzbogaconego estru.[8] Ostatnią z omawianych strategii jest kataliza

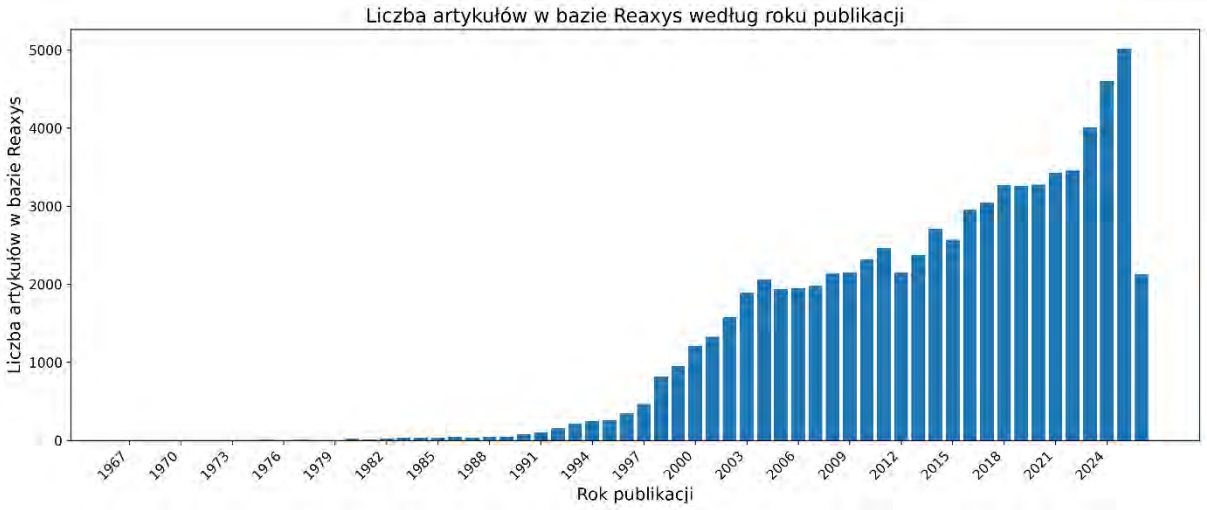
asymetryczna, jedna z najważniejszych współczesnych metod otrzymywania związków enancjomerycznie wzbogaconych, w której chiralny katalizator sprzyja powstawaniu jednego z enancjomerów w nadmiarze. Przykładem asymetrycznej katalizy w syntezie jest enancjoselektywne uwodornienie prochiralnego enamidu w metodzie otrzymywania leku sitagliptyny. Katalizowane jest kompleksem rodu z chiralnym ligandem Josiphos.[9] Produkt w postaci soli fosforowej sitagliptyny został otrzymany z wysoką wydajnością rzędu 96% , czystością >99.9% oraz „*nearly perfect ee*".



Rysunek 1. *Strategie otrzymywania związków enancjomerycznie wzbogaconych.*

W XXI wieku szczególnie intensywnie rozwijana jest kataliza asymetryczna, umożliwiająca prowadzenie syntezy enancjoselektywnej przy użyciu katalitycznych ilości chiralnych związków organicznych (organokataliza) lub kompleksów metali (kataliza homogeniczna i heterogeniczna). W przeciwieństwie do metod wykorzystujących pulę chiralną, pomocniki chiralne czy rozdział kinetyczny podejście to pozwala ograniczyć straty materiałowe oraz zwiększyć ekonomię atomową procesu.[10] Jedna cząsteczka chiralnego katalizatora może teoretycznie prowadzić do powstania dużej liczby cząsteczek produktu optycznie czynnego, co czyni katalizę asymetryczną szczególnie atrakcyjną z punktu widzenia efektywności syntetycznej.

1.3 Kataliza asymetryczna i skala energetyczna problemu



***Rysunek 2**. Liczba artykułów dostępnych w bazie Reaxys dla słów kluczy 'asymmetric' or 'chiral' and 'catalysis'.*

Rosnące zainteresowanie katalizą asymetryczną znajduje odzwierciedlenie w liczbie publikacji zebranych w bazie danych Reaxys (Rysunek 2). Tylko w 2025 r. opublikowano prawie 5000 artykułów pod hasłem „kataliza asymetryczna" (na podstawie wyników wyszukiwania w maju 2026).

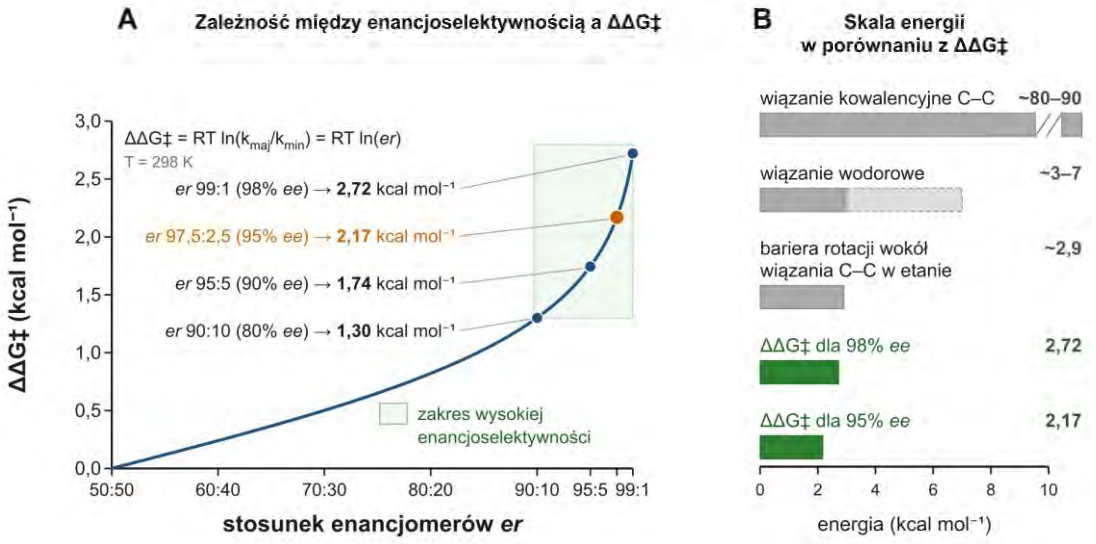
Znaczenie katalizy asymetrycznej dla współczesnej chemii potwierdzają dwie Nagrody Nobla. W 2001 roku otrzymali ją W. S. Knowles i R. Noyori za prace nad chiralnie katalizowanymi reakcjami uwodornienia oraz K. B. Sharpless za prace nad chiralnie katalizowanymi reakcjami utleniania,[11] a w 2021 roku B. List i D. W. C. MacMillan za rozwój asymetrycznej organokatalizy.[12]

Jednak nawet przy dostępie do odpowiednio dużej liczby reakcji i wyników dla wybranych metod katalitycznych świadome projektowanie katalizatorów pozostaje jednym z największych wyzwań w syntezie asymetrycznej. Trudność projektowania takich układów wynika między innymi z bardzo małej skali energetycznej odpowiadającej obserwowanej enancjoselektywności.[13] Różnica energii swobodnej aktywacji pomiędzy dwiema diastereomerycznymi strukturami przejściowymi (transition states), prowadzącymi do przeciwnych enancjomerów produktu, jest niewielka w porównaniu z energiami typowych wiązań chemicznych, co przedstawiono porównawczo na Rysunku 3B. Przy założeniu kontroli kinetycznej zależność między stosunkiem

szybkości powstawania obu enancjomerów a różnicą energii swobodnej aktywacji można zapisać jako:

$$| \Delta\Delta G^{\ddagger} | = RT \ln\left(\frac{k_{\text{major}}}{k_{\text{minor}}}\right) \approx RT\ln(er)$$

Zależność tę przedstawiono na Rysunku 3A. Dla stosunku enancjomerów 97,5:2,5, czyli 95% *ee*, w temperaturze 25 °C odpowiada to różnicy energii swobodnej aktywacji wynoszącej około 2,17 kcal mol⁻¹. Jak pokazano na Rysunku 3B, jest to wartość niewielka w porównaniu z typowymi energiami wiązań chemicznych i zbliżona do innych małych skal energetycznych spotykanych w chemii. Oznacza to, że ze stosunkowo niewielkich zmian energetycznych pomiędzy konkurencyjnymi drogami reakcji może wynikać duża zmiana enancjoselektywności.



***Rysunek 3**. Zależność między enancjoselektywnością a różnicą energii swobodnej aktywacji oraz porównanie z wybranymi skalami energetycznymi w chemii. (A) Zależność między stosunkiem enancjomerów a wartością |ΔΔG‡| w temperaturze 25 °C, obliczoną przy założeniu kontroli kinetycznej. (B) Porównanie wartości |ΔΔG‡| odpowiadających wysokiej enancjoselektywności z wybranymi wielkościami energetycznymi charakterystycznymi dla procesów i oddziaływań chemicznych. Wartości przedstawione w panelu B mają charakter orientacyjny i służą porównaniu skali energetycznej.[13]*

Zależność ta ma dwie konsekwencje, do których wracam w całej pracy. Po pierwsze, na tak niewielkie różnice mogą wpływać jednocześnie efekty steryczne, elektronowe i konformacyjne, oddziaływania katalizatora z substratem, solwatacja oraz temperatura. Tak wiele czynników wpływających na wynik reakcji trudno rozdzielić w opisie ilościowym. Po drugie, wielkość, którą należy odtworzyć, mieści się w zakresie zaledwie kilku kcal mol⁻¹, dlatego nawet niewielkie błędy w obliczonych energiach mogą istotnie wpływać na przewidywaną enancjoselektywność.[13] To sprawia, że zgodność obliczeń z eksperymentem jest mniej oczywista, niż sugerowałaby dojrzałość samych metod.

1.4 Jak projektuje się katalizatory

Mimo rozwoju samej katalizy asymetrycznej strategia projektowania nowych katalizatorów pozostawała przez długi czas zaskakująco niezmienna. Dominującym podejściem jest nadal

eksperymentalny przesiew bibliotek ligandów oraz ich późniejsza modyfikacja oparta na intuicji chemicznej i wiedzy mechanistycznej. Ten empiryczny proces ma charakter iteracyjny: identyfikuje się obiecujące struktury na podstawie doświadczenia, analizuje wpływ podstawników oraz efektów sterycznych i elektronowych także za pomocą doświadczeń, a następnie projektuje kolejne generacje ligandów – tylko jeżeli wyniki doświadczeń rysują interpretowalne zależności. Podejście to jest niezwykle czasochłonne i materiałochłonne, szczególnie dla nowoczesnych układów katalitycznych obejmujących dużą liczbę potencjalnych parametrów strukturalnych.[14]

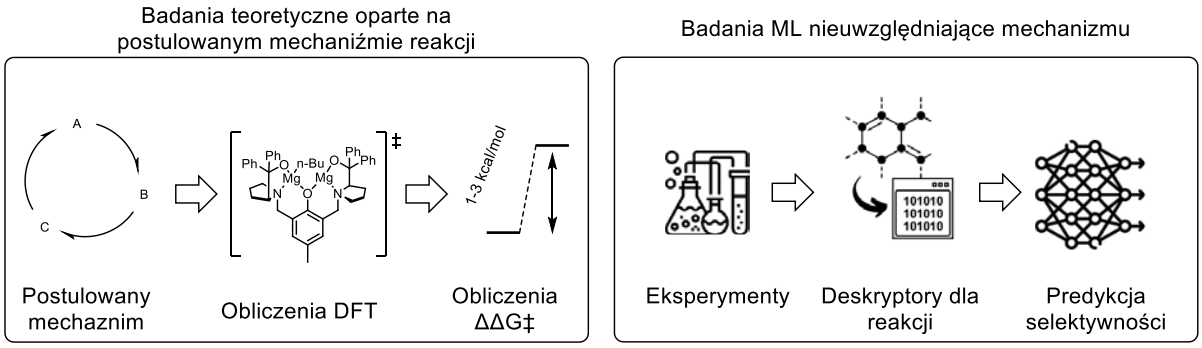
Ograniczeniem klasycznych metod jest również trudność ilościowego opisu zależności pomiędzy strukturą katalizatora a selektywnością reakcji. Historycznie próby takiego opisu opierały się na stałych Hammetta oraz na deskryptorach sterycznych i elektronowych wyznaczanych eksperymentalnie lub metodami chemii kwantowej.[15, 16] Podejścia te znalazły zastosowanie w modelach typu QSPR, QSRR i QSSR, jednak ich skuteczność zależała zwykle od wcześniejszego, intuicyjnego wyboru odpowiednich deskryptorów przez chemika. W praktyce oznaczało to, że modele budowano dopiero po wstępnej optymalizacji układu katalitycznego, gdy najważniejsze cechy strukturalne były już częściowo zidentyfikowane eksperymentalnie, a więc wtedy, gdy ich wartość predykcyjna była najmniejsza. Modele oraz hipotezy mechanistyczne buduje się więc najczęściej „wstecz": dopiero po przeprowadzeniu serii eksperymentów optymalizacyjnych i zaobserwowaniu zależności pomiędzy strukturą katalizatora, substratu lub warunkami reakcji a uzyskanym wynikiem.

1.5 Dlaczego wyznaczanie enancjoselektywności drogą obliczeniową jest wyzwaniem?

Selektywność reakcji asymetrycznej da się w zasadzie wyznaczyć metodą obliczeniową: wystarczy wyznaczyć struktury stanów przejściowych prowadzących do obu enancjomerów i porównać ich energie. W praktyce droga ta jest jednak zarówno niepewna, jak i kosztowna. Wymaga postawienia hipotezy mechanistycznej, wskazania etapu determinującego stereoselektywność, przeszukania przestrzeni konformacyjnej dla każdej rozważanej struktury i obliczeń kwantowochemicznych na poziomie wystarczającym, by różnice rzędu jednej kilokalorii na mol były wiarygodne. Nakład ten trzeba ponieść od nowa dla każdego nowego katalizatora i każdej nowej reakcji, a i tak nie mamy gwarancji, że cały trud się opłaci, ponieważ przyjęta przez nas hipoteza mechanistyczna może być błędna.

Podejście oparte na danych odwraca ten porządek: zamiast wyprowadzać selektywność z mechanizmu, poszukuje się zależności między obliczalnymi cechami układu a wynikiem zmierzonym doświadczalnie. Skoro ryzyko wyznaczenia niepoprawnego mechanizmu reakcji jest tak duże to atrakcyjną alternatywą jest podejście, które nie wymaga hipotezy mechanistycznej, lecz wymaga danych - i to danych opisanych w sposób jednorodny. Obie drogi zestawia Rysunek 4.

Wybór między nimi nie jest wyborem między dokładnością a wygodą, lecz wyborem pomiędzy nakładem pracy obliczeniowej a nakładem pracy przy tworzeniu jakościowej bazy danych. Nie mogę i nie postuluję, że któraś z dróg jest znacząco lepsza, ponieważ obie mają kluczowy wpływ na poznawanie arkanów katalizy asymetrycznej. Moja rozprawa doktorska dotyczy drogi drugiej, a znaczna część opisanej w niej pracy przypada właśnie na opracowanie danych.



Rysunek 4. *Dwie drogi przewidywania selektywności reakcji asymetrycznej. Po lewej: podejście oparte na postulowanym mechanizmie — cykl katalityczny wyznacza strukturę przejściową, którą optymalizuje się metodą DFT, a różnica energii swobodnej aktywacji ΔΔG‡ rzędu 1–3 kcal mol⁻¹ przekłada się bezpośrednio na obserwowany nadmiar enancjomeryczny (Rysunek 3). Po prawej: podejście oparte na danych — wyniki eksperymentalne opisuje się deskryptorami reakcji, a model uczenia maszynowego przewiduje selektywność bez odwoływania się do mechanizmu.*

2. Kataliza asymetryczna z wykorzystaniem kompleksów magnezu

2.1 Właściwości kompleksów magnezu istotne dla katalizy

Jon Mg^{2+} jest kationem o małym promieniu jonowym i wysokiej gęstości ładunku, przez co wykazuje wyraźną oksofilowość i wiąże przede wszystkim donory tlenowe i azotowe: grupy karbonylowe, alkoholanowe, fenolanowe oraz iminowe. Jako kwas Lewisa aktywuje elektrofil przez koordynację do atomu tlenu grupy karbonylowej, obniżając energię orbitalu LUMO i zwiększając podatność węgla karbonylowego lub węgla w sprzężonym, elektrofilowym wiązaniu podwójnym na atak nukleofilowy. Liczba koordynacyjna magnezu w typowych układach katalitycznych mieści się w zakresie od czterech do sześciu, co sprzyja wykorzystaniu go w katalizie, gdzie istotna jest koordynacja chiralnego liganda oraz substratu/substratów podczas cyklu katalitycznego.

Cechą, która odróżnia magnez od metali szlachetnych stosowanych w katalizie asymetrycznej, jest labilność sfery koordynacyjnej. Wymiana ligandów zachodzi szybko, co sprzyja obrotowi katalitycznemu, ale jednocześnie utrudnia otrzymanie jednej, dobrze zdefiniowanej formy katalitycznej. W praktyce oznacza to, że o wyniku reakcji decyduje nie tylko struktura liganda chiralnego, lecz również prekursor magnezowy, z którego katalizator generuje się *in situ*.

Osobną możliwością, którą otwiera chemia magnezu, jest jego dualna natura w wybranych związkach kompleksowych. Alkoholany i fenolany magnezu mogą działać jako zasada Brønsteda, deprotonując nukleofil, podczas gdy centrum metaliczne aktywuje elektrofil jako kwas Lewisa. Obaj partnerzy reakcji zostają wówczas związani w obrębie jednej sfery koordynacyjnej, a stereochemia produktu jest determinowana przez geometrię tak powstałego adduktu. Koncepcję tę wprowadzono dla kompleksów metali z ligandami typu BINOL,[17] natomiast w przypadku układów magnezowych jej zasadność wspierały badania zespołu Lewińskiego nad strukturą alkoksydów i alkiloperoksydów magnezu, które wskazywały na znaczenie struktur zawierających więcej niż jedno centrum metaliczne.[18]

2.2 Zakres transformacji i klasy chiralnych ligandów

Kataliza asymetryczna kompleksami magnezu obejmuje szeroki zakres przemian tworzących wiązania węgiel–węgiel, węgiel–tlen, węgiel–azot i węgiel–siarka.[19] Skalę tej chemii dobrze oddaje liczebność bazy danych zbudowanej na potrzeby pracy opisanej w rozdziale dotyczącym badań własnych.

Zakres transformacji asymetrycznych katalizowanych magnezem porządkuje przegląd Yanga i współpracowników, który wyróżnia dwie strategie generowania chiralnych kompleksów Mg(II).[19] W pierwszej z nich sól magnezu o wyraźnym charakterze kwasu Lewisa - $Mg(OTf)_2$, $Mg(NTf_2)_2$, $Mg(ClO_4)_2$ albo MgI_2 - koordynuje ligand chiralny, dając katalizator działający jako chiralny kwas Lewisa. W drugiej katalizator powstaje *in situ* w reakcji zobojętnienia z użyciem prekursorów o wyraźniejszej zasadowości, takich jak alkoholany $Mg(OR)_2$ lub dialkilomagnez MgR_2. Ligand chiralny pełni wówczas jednocześnie funkcję przeciwjonu, a powstający kompleks łączy kwasowość Lewisa z zasadowością Brønsteda.

Wśród ligandów chiralnych stosowanych z magnezem powtarza się kilka klas. Do najczęściej spotykanych należą pochodne BINOL-u oraz ich pochodne w postaci chiralnych kwasów fosforowych (CPA), ligandy bis(oksazolinowe), zasady Schiffa typu salen oraz aminoalkohole i diaminy. Osobną grupę stanowią ligandy tworzące kompleksy bicentrowe, do których należy omawiany dalej ProPhenol. Rozpiętość tych rodzin jest jednym z powodów, dla których wybór liganda dla nowej reakcji magnezowej pozostaje nieoczywisty: nie istnieje jeszcze reguła wiążąca typ przemiany z rodziną liganda, a przesłanki są rozproszone po dziesiątkach publikacji, z których każda opisuje własny, wąski zakres substratowy.

Ta rozproszona i niejednorodna literatura sprawia jednocześnie, że kataliza magnezowa dobrze nadaje się do analizy opartej na danych. Wyniki zebrane z kilkuset prac obejmują wystarczająco szeroki zakres reakcji i układów katalitycznych, aby poszukiwać zależności pomiędzy rodzajem reakcji a zastosowanym katalizatorem, pod warunkiem wcześniejszego uporządkowania i ujednolicenia danych.

Przegląd Czombik, Gajewy, Czapik i Kwita zwraca dodatkowo uwagę na aspekt strukturalny - wyznaczoną doświadczalnie geometrię form katalitycznych oraz obliczone struktury stanów przejściowych. Podkreślając przy tym, że w części układów zależność pomiędzy strukturą kompleksu a obserwowaną stereoselektywnością została szczegółowo przeanalizowana, jednak wiedza ta dotyczy pojedynczych klas reakcji i katalizatorów. Autorzy wskazują, że mimo obfitości magnezu w skorupie ziemskiej liczba reakcji asymetrycznych, w których pełni on rolę katalizatora, pozostaje niewielka w porównaniu z katalizą metalami przejściowymi i organokatalizą.[20]

2.3 Magnez jako atrakcyjny metal w katalizie i syntezie

Poza względami chemicznymi za magnezem przemawiają przesłanki praktyczne, które warto sformułować ostrożnie. Pallad, iryd, rod, ruten i osm należą do platynowców: pierwiastków o niskiej zawartości w skorupie ziemskiej, silnie skoncentrowanej geograficznie podaży i dużej zmienności cen.[21, 22, 23] Magnez jest natomiast jednym z pierwiastków pospolitych, a jego związki są dostępne w skali technicznej.[20]

W wymiarze regulacyjnym wytyczna ICH (Międzynarodowa Rada ds. Harmonizacji, ang. *International Council for Harmonisation*) Q3D(R2) zalicza pallad, iryd, rod i ruten do klasy 2B i

wyznacza dla nich wartości dopuszczalnego dziennego narażenia zależne od drogi podania.[24] Dla podania doustnego wartość ta wynosi 100 µg na dobę zarówno dla palladu, jak i dla irydu, rodu i rutenu (dla podania pozajelitowego 10 µg na dobę), przy czym dla irydu, osmu, rodu i rutenu wytyczna przyjęła wartości ustalone dla palladu z powodu niewystarczających danych toksykologicznych.[24] Magnez nie jest klasyfikowany przez wytyczne jako zanieczyszczenie pierwiastkowe budzące obawy i nie wyznacza dla niego takiej wartości.[24] Brak wyznaczonej wartości oznacza, że pierwiastek pozostaje poza zakresem tej wytycznej, a nie że jego dowolny związek jest bezpieczny. Fizjologiczna obecność magnezu w organizmie nie stanowi przesłanki, że każdy wykorzystywany związek kompleksowy magnezu nie stanowi dla niego zagrożenia.

Proces usunięcia pozostałości metalu szlachetnego z substancji czynnej jest osobnym zadaniem technologicznym, którego koszt rośnie wraz z ostrością limitu. Niestety kataliza magnezem nie może być uznawana od razu za atrakcyjny zamiennik katalizy opartej na metalach szlachetnych – różnice we właściwościach elektronowych będą stały na przeszkodzie katalizowania tych samych reakcji - ale nie w każdym przypadku. Odpowiednio zaprojektowane otoczenie wokół centrum metalicznego może jednak w znacznym stopniu modyfikować reaktywność kompleksu i umożliwiać mu udział w przemianach, które są trudne dla prostszych związków tego metalu. Przykładem są kompleksy magnezu z ligandami kleszczowymi typu PNN, w których kooperacja metal–ligand, oparta na odwracalnej aromatyzacji i dearomatyzacji szkieletu liganda, umożliwiła aktywację wiązania H–H i katalityczne uwodornienie imin oraz *N*-heteroarenów [25]. Autorzy wykazali przy tym, że dearomatyzowana forma kompleksu była kluczowa dla aktywacji H_2 i wysokiej aktywności katalitycznej.

Zestawienie tych przesłanek z materiałem omówionym w rozdziale 3 prowadzi do obserwacji, od której zaczyna się część badawcza rozprawy: kataliza magnezowa dysponuje literaturą dostatecznie obszerną, by budować na niej modele predykcyjne.

3. Jak buduje się i ocenia model predykcyjny – kompendium do zrozumienia rozprawy

3.1 Od zapisu reakcji do wektora liczbowego

W chemii komputerowej reprezentacja danych, czyli deskryptory, oznacza sposób przełożenia struktury cząsteczki, układu reakcyjnego lub warunków procesu na postać możliwą do przetwarzania przez model matematyczny. W praktyce najczęściej przyjmuje ona formę deskryptorów molekularnych liczbowych lub binarnych, gdzie każdy opisuje wybrane cechy układu chemicznego. Transformacja ta jest konieczna, ponieważ klasyczne modele uczenia maszynowego operują na wektorach cech, a więc wymagają danych zapisanych w postaci liczb.

Deskryptory nie są jednorodną grupą - mogą opisywać proste właściwości globalne, topologię cząsteczki, jej geometrię przestrzenną lub bardziej wyspecjalizowane cechy zaprojektowane pod konkretny problem chemiczny. Intuicyjna systematyzacja przewiduje podział według rodzaju kodowanej informacji i wynikającego z niej kompromisu między zawartością informacji chemicznej a kosztem obliczeniowym. Generowanie deskryptorów stanowi pomost między intuicją chemiczną a modelowaniem opartym na danych.[26] Chemik patrzący na cząsteczkę rozpoznaje grupy funkcyjne, efekty elektronowe i zawadę przestrzenną, a dla algorytmu pojęcia te nie istnieją, dopóki nie zostaną wyrażone liczbą. Jakość modelu zależy więc przede

wszystkim od tego, czy wśród deskryptorów molekularnych naszych reakcji znajdują się te reprezentacje cech mające wpływ na interesującą nas zależność.

Fingerprinty i deskryptory fragmentaryczne

Szczególną klasą deskryptorów są fingerprinty molekularne - wektorowe reprezentacje struktury cząsteczki (ang. *Fingerprint* – odcisk palca). Najczęściej mają postać wektorów binarnych, w których poszczególne pozycje informują o obecności lub braku określonych motywów strukturalnych, a w niektórych wariantach zamiast informacji binarnej stosuje się częstość występowania danego fragmentu.

Szeroko stosowanymi fingerprintami są fingerprinty Morgana, znane również jako ECFP (*extended-connectivity fingerprints*). Są to fingerprinty cyrkularne: dla każdego atomu analizowane jest jego lokalne otoczenie w zadanym promieniu, a zakodowane fragmenty mapowane są następnie na wektor cech. Zmiana promienia kontroluje zakres uwzględnianej informacji strukturalnej — mniejszy promień podkreśla lokalne środowisko atomowe, większy pozwala uchwycić bardziej rozległe motywy strukturalne. Obok fingerprintów Morgana stosowane są również inne reprezentacje, takie jak RDKit fingerprints czy Avalon fingerprints, różniące się sposobem definiowania i kodowania fragmentów.[27, 28, 29]

W przypadku struktur bogatych w skondensowane układy pierścieniowe klasyczne podejścia fragmentaryczne mogą okazać się niewystarczające. Z tego względu opracowano deskryptory CircuS (*Circular Substructures*),[14] które jawnie uwzględniają pełne podstruktury zawierające zamknięte pierścienie, co pozwala lepiej opisać motywy cykliczne i policykliczne. Uzupełnienie stanowią deskryptory ChyLine, oparte na liniowych fragmentach o zadanej długości, zliczanych bezpośrednio z grafu cząsteczki.[30]

Osobną kategorią są biblioteki deskryptorów obliczanych masowo dla całych zbiorów cząsteczek. Do często wykorzystywanych narzędzi należą RDKit i Mordred, przy czym Mordred umożliwia obliczenie ponad tysiąca deskryptorów dwuwymiarowych i trójwymiarowych.[28, 31] Obejmują one między innymi indeksy topologiczne, deskryptory autokorelacyjne typu Moreau–Broto i Geary'ego ważone właściwościami atomowymi, miary złożoności struktury oraz proste deskryptory zliczające.[31] Ich zaletą jest stosunkowo niski koszt obliczeniowy, szczególnie w przypadku deskryptorów 2D, oraz możliwość stosowania do cząsteczek o zróżnicowanej budowie. Wadą może być natomiast duża liczba cech w stosunku do niewielkiej liczby obserwacji, co w małych zbiorach zwiększa ryzyko przeuczenia i wymaga odpowiedniej selekcji lub redukcji liczby cech albo zastosowania metod regularyzacji.

Obok deskryptorów strukturalnych do reprezentacji włącza się niekiedy proste wielkości elektronowe wyprowadzane z cząstkowych ładunków atomowych, na przykład ładunków Gasteigera, oraz parametry fizykochemiczne rozpuszczalnika i warunków prowadzenia reakcji. Wielkości te nie wymagają obliczeń kwantowochemicznych, a wnoszą informację, której deskryptory topologiczne nie zawierają.

Deskryptory reakcji

Reprezentacja danych może dotyczyć również całych reakcji chemicznych, a nie jedynie pojedynczych cząsteczek. W tym celu stosuje się m.in. *Condensed Graph of Reaction* (CGR) - reprezentację, w której substraty i produkty łączone są w jeden graf pseudocząsteczkowy zawierający informację o wiązaniach powstających, zrywanych lub zmieniających swój charakter

w trakcie przemiany. Dzięki temu klasyczne deskryptory 2D, w tym fingerprinty i deskryptory fragmentaryczne, mogą być stosowane nie do osobnych reagentów, lecz do reakcji jako całości.

Różnicowy fingerprint reakcji (DRFP) realizuje tę samą myśl prościej. Generuje się fingerprinty dla substratów i produktów, a następnie zapisuje ich różnicę symetryczną wektora fingerprintu. Otrzymany wektor binarny opisuje wyłącznie zmianę strukturalną, nie zaś to, co w reakcji pozostało nienaruszone.

Deskryptory „szyte na miarę"

Obok uniwersalnych reprezentacji dostępnych w bibliotekach chemoinformatycznych istotną rolę odgrywają deskryptory tworzone specjalnie pod określony problem badawczy. Jest to szczególnie ważne przy opracowywaniu ilościowych zależności struktura–reaktywność, gdy różnice steryczne i elektronowe w bezpośrednim otoczeniu centrum katalitycznego są na tyle subtelne, że dostępne metody nie zawsze pozwalają je uchwycić. Rośnie też zapotrzebowanie na nowe sposoby opisu reakcji w katalizie asymetrycznej, szczególnie w przypadku kompleksów metali.[32]

Do klasycznych przykładów należą parametry Sterimol,[33, 34] opisujące rozmiar podstawnika względem przyjętej osi: parametr L odpowiada jego długości, natomiast B_1 i B_5 opisują odpowiednio minimalny i maksymalny wymiar poprzeczny. Pokrewny charakter ma stosowany w chemii kompleksów metali kąt Tolmana, opisujący rozwarcie kątowe stożka zajmowanego przez ligand oraz procentowa objętość zajęta przez ligand ($\%V_{bur}$, buried volume), określająca, jaka część umownej sfery otaczającej centrum metaliczne jest przestrzennie wypełniona przez ligand. [35, 36]

Przykładem bardziej zaawansowanego i ciekawego podejścia z roku 2022 są deskryptory SMART (*Spatial Molding for Approachable Rigid Targets*), opracowane przez grupę Sigmana do ilościowego opisu objętości i kształtu strefy koordynacyjnej w katalizatorach rodu.[37] W metodzie tej do centrum katalitycznego wprowadza się sztuczną sondę molekularną — makrocykl siarkowy o dużej swobodzie konformacyjnej — a następnie przeszukuje jej możliwe konformacje tak, aby odwzorować przestrzeń rzeczywiście dostępną dla koordynującego substratu. Metoda ta pozwala wyznaczyć parametry opisujące objętość kieszeni koordynacyjnej oraz jej kształt i dostępność steryczną [37].

Deskryptory eksperckie tego typu pokazują, że reprezentacja danych może być projektowana świadomie i celowo, a nie jedynie wybierana z katalogu gotowych opcji. W wielu zastosowaniach okazują się one bardziej informatywne niż ogólne fingerprinty, ponieważ kodują bezpośrednio cechy chemicznie istotne dla badanego mechanizmu lub selektywności. Jednak nawet tutaj istnieje ryzyko, że badacz, wybierając z góry określone cechy, nadmiernie zawęzi obszar poszukiwania zależności przez model, pomijając właściwości układu, które rzeczywiście mają znaczenie.

Reprezentacje językowe

Odrębną drogą jest potraktowanie zapisu SMILES jako sekwencji znaków i wyuczenie reprezentacji cząsteczki tak, jak uczy się reprezentacji słów w przetwarzaniu języka naturalnego. Modele z rodziny ChemBERTa, wstępnie trenowane na dużych zbiorach struktur, zwracają dla każdej cząsteczki wektor, którego poszczególne współrzędne nie mają interpretacji chemicznej, ale który koduje podobieństwo strukturalne wyuczone z danych. Zaletą jest brak konieczności projektowania deskryptorów, wadą - wysoka wymiarowość i brak interpretowalności.

Podsumowując, w chemii komputerowej nie istnieje jedna uniwersalna reprezentacja danych najlepsza dla wszystkich zastosowań. Ten sam układ może być opisany różnymi reprezentacjami, a ich przydatność jest zadaniowo zależna. Reprezentacje można wybierać spośród gotowych rozwiązań, jak również konstruować samodzielnie tak, aby odpowiadały specyfice badanego układu. Dobór reprezentacji stanowi jeden z kluczowych etapów budowy modelu predykcyjnego, ponieważ decyduje o tym, jaka informacja chemiczna zostaje zachowana, a jaka utracona.

Wybór między reprezentacją opartą na deskryptorach o jasnej interpretacji fizycznej a reprezentacją fragmentową lub uczoną nie jest przy tym neutralny. Deskryptory kwantowochemiczne są kosztowne, ale interpretowalne i skuteczne przy małej liczbie obserwacji, natomiast fingerprinty i sieci grafowe pozwalają opisać miliony struktur w rozsądnym czasie, lecz wymagają większych zbiorów i tracą przejrzystość. Z tego powodu w wytycznych metodycznych dla katalizy homogenicznej zaleca się reprezentacje fizyczne dla małych zbiorów danych o wysokiej jakości, a reprezentacje fragmentowe i uczone rezerwuje się dla danych wielkoskalowych.[26]

Zalecenie to może stanowić punkt odniesienia, ale nie jest regułą. Tsuji i współpracownicy wykazali, że deskryptory fragmentowe dorównują siatkowym deskryptorom trójwymiarowym na tym samym zbiorze katalizatorów, przy zerowym koszcie kwantowochemicznym.[14]

3.2 Co należy modelować: ΔΔG‡ czy %*ee*

Nadmiar enancjomeryczny jest wielkością ograniczoną i nieliniowo związaną z różnicą energii swobodnej aktywacji, co omówiono w podrozdziale 1.3. Regresja prowadzona bezpośrednio na %*ee* nie uwzględnia nieliniowej zależności między enancjoselektywnością a ΔΔG‡, dlatego taka sama różnica w wartości *ee* może odpowiadać różnym różnicom energetycznym w zależności od miejsca na skali. Dlatego często modelowaną zmienną jest ΔΔG‡. Przykładowo przejście od 90 do 95% *ee* odpowiada zmianie o około 0,43 kcal mol^{-1}, a od 95 do 99% *ee* — o około 0,97 kcal mol^{-1}.

ΔΔG‡ nie zawsze jest jednak najlepszym wyborem dla danych literaturowych, ponieważ jej wyznaczenie z *ee* wymaga znajomości temperatury, w której powstaje produkt z danym nadmiarem enancjomerycznym. Jeżeli reakcję prowadzono w gradiencie temperatury, nie wiadomo, w jakiej temperaturze przereagował konkretny substrat — bardziej reaktywne związki mogą reagować już w niższej temperaturze, a mniej reaktywne dopiero po jej podwyższeniu. W takim przypadku eksperymentalne *ee* pozostaje dobrze określoną wartością, natomiast przypisanie stałej wartości temperatury do obliczenia ΔΔG‡ staje się niejednoznaczne.

3.3 Rozkład danych i zakres stosowalności

3.3.1 Rozkład danych a wybór modelu

Ważne jest, jak rozłożone są wartości, których model ma się nauczyć. W regresji zbiór powinien obejmować możliwie szeroki zakres zmiennej odpowiedzi, a w klasyfikacji należy zwracać uwagę na liczebność poszczególnych klas.[26] Jeżeli większość wyników skupia się w jednym zakresie, model ma mało informacji o pozostałych przypadkach i może osiągać dobre metryki głównie dzięki poprawnemu przewidywaniu wartości najczęstszych.

W katalizie asymetrycznej jest to szczególnie istotne, ponieważ literatura jest przesunięta w stronę wyników udanych. Wyniki o niskiej selektywności lub wydajności oraz reakcje nieudane są publikowane rzadziej. Nie można jednak po prostu wyrównać rozkładu przez usunięcie części dobrych wyników, ponieważ oznaczałoby to utratę dostępnych danych. Lepszym rozwiązaniem jest odzyskiwanie mniej udanych rezultatów z badań optymalizacyjnych i materiałów uzupełniających.

Zebrana przeze mnie baza reakcji katalizowanych magnezem dobrze pokazuje ten problem. Ręczna analiza materiałów uzupełniających pozwoliła włączyć 128 reakcji o nadmiarze enancjomerycznym poniżej 50%, mimo czego rozkład *ee* nadal pozostawał wyraźnie przesunięty ku wysokim wartościom.[38] Pokazuje to, że wyniki mniej udane są szczególnie cenną częścią zbioru, ponieważ dostarczają modelowi informacji o granicach działania badanego układu.

3.3.2 Zakres stosowalności

Model odpowiada na każde pytanie, także takie, na które nie ma podstaw odpowiadać. Zakresem stosowalności nazywa się obszar przestrzeni chemicznej, w którym predykcja jest poparta danymi treningowymi. W praktyce wyznacza się go przez odległość obiektu od zbioru treningowego w przestrzeni deskryptorów albo przez maksymalne podobieństwo do obiektów treningowych.

Odpowiednikiem chemicznym jest tabela zakresu substratowego: publikując metodę, podaje się substraty, dla których ją sprawdzono, i nie twierdzi się, że działa dla wszystkich. Model bez zadeklarowanego zakresu stosowalności jest odpowiednikiem metody bez tabeli zakresu działania tej metody na wybranych substratach.

3.4 Podział danych i walidacja

Sposób podziału danych powinien uwzględniać nie tylko ich liczebność, lecz także wzajemne podobieństwo badanych obiektów. Jeżeli zbiór składa się z kilku grup bardzo podobnych związków, losowy podział może umieścić bliskie analogi zarówno w zbiorze treningowym, jak i testowym. Ocena modelu będzie wtedy odpowiadała przede wszystkim na pytanie, czy potrafi przewidywać wyniki dla związków podobnych do tych, które już widział.

3.4.1 Po co dzieli się zbiór

Model można bardzo dobrze dopasować do danych, na których został zbudowany. Znacznie ważniejsze jest jednak pytanie, czy poprawnie przewidzi wynik dla nowych danych i potrafi generalizować.

Zbiór dzieli się zazwyczaj na część treningową, wykorzystywaną do dopasowania modelu, oraz dane niewykorzystywane bezpośrednio do uczenia. W zależności od przyjętej procedury wydziela się dodatkowo zbiór walidacyjny, służący do wyboru modelu i jego parametrów, oraz niezależny zbiór testowy, wykorzystywany do końcowej oceny skuteczności.

Selekcję cech również należy prowadzić wyłącznie z wykorzystaniem danych treningowych. Jeżeli wyboru cech dokonuje się na całym zbiorze przed jego podziałem, informacja

o wynikach ze zbioru testowego pośrednio wpływa na budowę modelu, prowadząc do zawyżenia jego ocenianej skuteczności. Zasadę tę stosowaliśmy w badaniach opisanych w rozdziale 6.2.

3.4.2 Walidacja krzyżowa

Przy stosunkowo małych zbiorach, a takie są w katalizie asymetrycznej większością, wydzielenie stałego wewnętrznego zbioru testowego jest kosztowne, ponieważ odbiera modelowi część i tak nielicznych obserwacji. Stosuje się wtedy walidację krzyżową: zbiór dzieli się na k równych części, model trenuje się k razy, za każdym razem pomijając jedną z nich i na niej sprawdzając wynik. Wariantem skrajnym jest walidacja z pominięciem jednej reakcji (LORO – *Leave one reaction out*), w której k równa się liczbie reakcji. Daje ona oszacowanie obciążone optymistycznie i przy bardzo małych zbiorach bywa niestabilna.[26, 39] Wytyczne dobrych praktyk dla katalizy homogenicznej zalecają wprost walidację pięciokrotną.[26]

Osobno należy odróżnić walidację wewnętrzną od walidacji zewnętrznej, prowadzonej na danych, których model nie widział na żadnym etapie budowy. Walidacja zewnętrzna, a tym bardziej prospektywna, jest testem znacznie mocniejszym, a odpowiednio skonstruowane podziały grupowe badają natomiast określony, węższy rodzaj generalizacji i również niosą informację. Sposób podziału zbioru i wyboru części do walidacji zewnętrznej nie jest czynnością techniczną poprzedzającą modelowanie, lecz decyzją o tym, co właściwie zostanie zmierzone.

3.4.3 Punkty odniesienia

Wynik modelu jest liczbą bez znaczenia, dopóki nie zostanie z czymś porównany. Punktem odniesienia najprostszym jest model naiwny: w regresji model przypisujący wszystkim obiektom mediany lub średniej zmiennej odpowiedzi ze zbioru treningowego, w klasyfikacji jest to model wskazujący zawsze klasę najliczniejszą. Model, którego metryki nie przewyższają wyników modelu naiwnego, nie wykazuje rzeczywistej zdolności predykcyjnej.

Mocniejszym testem jest punkt odniesienia oparty na podobieństwie — model najbliższego sąsiada, przypisujący obiektowi testowemu wynik obiektu treningowego najbardziej do niego podobnego. Jeżeli model złożony nie wypada wyraźnie lepiej, to znaczy, że odtwarza wyłącznie podobieństwo strukturalne. Uzupełnieniem są kontrole na cechach nieinformatywnych: ten sam model zbudowany na cechach losowych albo na etykietach kodowanych one-hot, oraz przetasowanie zmiennej odpowiedzi, po którym model powinien stracić zdolność predykcyjną. Wynik nieodróżnialny od takiej kontroli oznacza, że dokładność pochodzi ze struktury podziału, a nie z chemii.

Dobrym przykładem znaczenia takich kontroli jest ponowna analiza danych Ahnemana i współpracowników dotyczących reakcji Buchwalda–Hartwiga. Chuang i Keiser zastąpili zastosowane deskryptory chemiczne losowymi wartościami, tworząc dla każdego składnika reakcji przypadkowy „kod”, a następnie powtórzyli modelowanie na tym samym podziale danych. Modele oparte na takich cechach osiągały wyniki niemal identyczne jak modele wykorzystujące rzeczywiste deskryptory chemiczne, co pokazało, że duża część zdolności predykcyjnej mogła wynikać ze struktury samego zbioru eksperymentalnego, a nie z uchwycenia zależności chemicznych.[40]

3.5 Metryki oceny modelu

W regresji wartość R^2 informuje, w jakim stopniu model odtwarza zróżnicowanie obserwowanych wartości, ale sama w sobie nie świadczy o jego zdolności predykcyjnej i powinna być interpretowana razem z wynikiem walidacji oraz miarami błędu. W praktycznych wytycznych dotyczących modeli MLR, wskazanych przez M. Sigmana i współpracowników, w chemii organicznej wartości R^2 poniżej około 0,6 wskazano jako zwykle świadczące o niewystarczającej jakości dopasowania, przy czym znaczenie tej wartości zależy od badanego problemu.[41] Kryterium użyteczności jest przy tym proste: błąd powinien mieścić się w granicach tolerancji chemika.

W klasyfikacji dokładność jest miarą intuicyjną, lecz przy niezbalansowanych klasach może zmylić, ponieważ właściwym punktem odniesienia nie jest zero, lecz udział klasy najliczniejszej. Odporniejsze są: ważona miara F1, łącząca precyzję z czułością z uwzględnieniem liczebności klas, oraz zbalansowana dokładność, będąca średnią czułości i swoistości, której wartość odniesienia wynosi 0,5 niezależnie od proporcji klas. Osobną miarą, użyteczną przy zadaniach rankingowych, jest dokładność top-n, podająca, czy właściwa klasa znalazła się wśród n pierwszych wskazań.

3.6 Interpretowalność modeli i analiza istotności cech

Cenną informacją po wykonaniu modelowania jest wskazanie, które cechy struktury cząsteczek wpływają na predykcję i w jakim kierunku. Jeżeli model wykorzystuje chemicznie interpretowalne deskryptory, wnioski można bezpośrednio odnieść do właściwości układu. Przykładowo, jeżeli wzrost $\%V_{bur}$ katalizatora, opisującego zatłoczenie steryczne wokół centrum metalicznego, jest w danym modelu związany ze wzrostem enancjoselektywności, może to sugerować, że zwiększenie osłony sterycznej tego centrum jest korzystnym kierunkiem modyfikacji struktury katalizatora.

Coraz częściej stosuje się jednak reprezentacje wysokowymiarowe, których poszczególne elementy nie mają tak bezpośredniego znaczenia chemicznego, np. fingerprinty lub reprezentacje wyuczone przez modele językowe. Ich wpływ na predykcję można analizować metodami interpretacji post hoc, takimi jak wartości SHAP (SHapley Additive exPlanations).[42] Wartości SHAP określają wkład poszczególnych cech w konkretną predykcję względem wartości bazowej modelu, a ich uśrednienie dla wielu przykładów pozwala ocenić, które cechy lub całe bloki informacji mają największe znaczenie dla działania modelu. W przypadku reprezentacji, których pojedynczych elementów nie można sensownie powiązać ze strukturą chemiczną, przydatna jest również analiza ablacyjna: model trenuje się ponownie po usunięciu wybranej grupy informacji, a spadek jego skuteczności pozwala ocenić, jak istotna była ona dla predykcji.

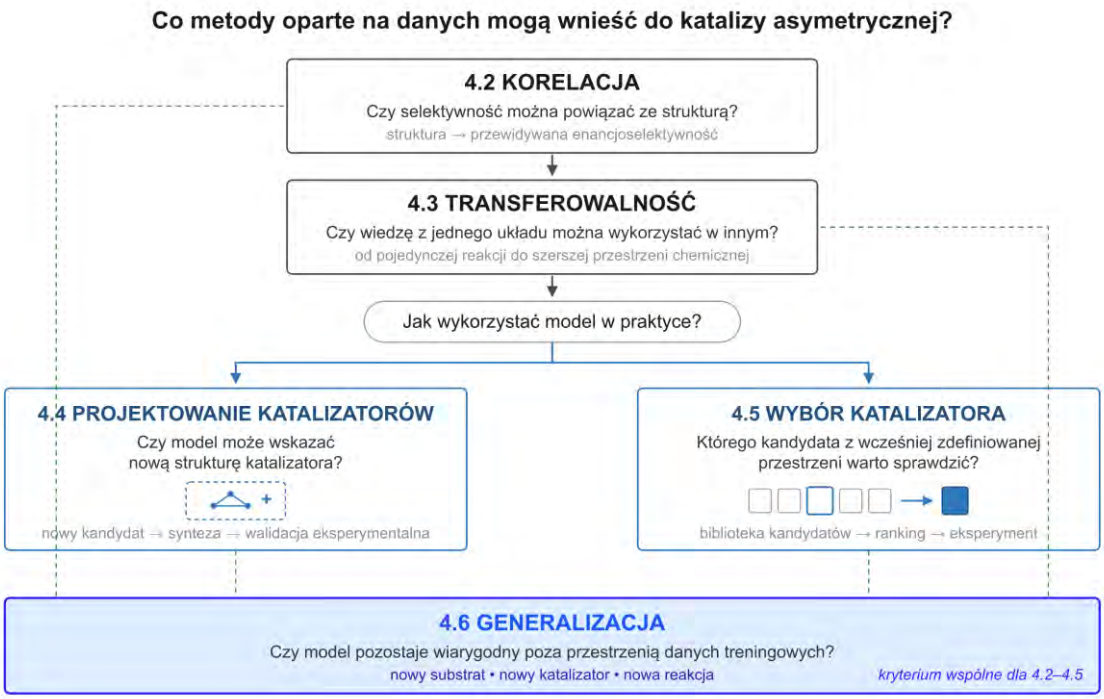
4. Uczenie maszynowe w katalizie asymetrycznej

4.1 Zakres rozdziału

W poniższym tekście omawiam prace, w których metody uczenia maszynowego i chemoinformatyki zastosowano do badania homogenicznej katalizy asymetrycznej. Poza zakresem pozostają kataliza heterogeniczna i biokataliza, a także badania dotyczące wyłącznie wydajności, regioselektywności lub optymalizacji warunków reakcji. Do takich prac odwołuję się jedynie wtedy, gdy stanowią istotny punkt odniesienia dla metod stosowanych w modelowaniu enancjoselektywności.

Metody oparte na danych rozwijano równolegle dla dwóch ważnych obszarów katalizy asymetrycznej: organokatalizy oraz homogenicznej katalizy kompleksami metali przejściowych. Wśród omawianych dalej przykładów dużą grupę stanowią reakcje organokatalityczne, szczególnie układy wykorzystujące chiralne kwasy fosforowe (CPA), dla których badano m.in. zależności struktura–selektywność, możliwość przenoszenia informacji pomiędzy reakcjami oraz uniwersalność katalizatorów. Równolegle rozwijano modele dla reakcji katalizowanych kompleksami palladu, miedzi czy rodu. W takich układach problem reprezentacji jest zwykle bardziej złożony, ponieważ poza strukturą substratów należy często uwzględnić również właściwości liganda, centrum metalicznego oraz geometrię powstającego kompleksu katalitycznego. Układy te cieszą się największą popularnością, ponieważ jednymi z najliczniej opisanych metod katalitycznych w literaturze.

Omawiane prace uporządkowano według pytania stawianego modelowi, a nie chronologicznie, co obrazuje Rysunek 5. Taki układ prowadzi od modeli opisujących zależności w obrębie określonego układu reakcyjnego do coraz trudniejszego problemu przewidywania dla nowych katalizatorów, reakcji i obszarów przestrzeni chemicznej.



Rysunek 5. *Mapa struktury Rozdziału 4. Strukturę uporządkowano według pytań, na które modele mogą odpowiadać w badaniach nad katalizą asymetryczną*

4.2 Korelacja — czy selektywność można zapisać jako funkcję struktury?

Punktem wyjścia jest pytanie typowo fizykoorganiczne: czy różnicę energii swobodnej aktywacji prowadzącą do powstania dwóch enancjomerów można powiązać ilościowo z mierzalnymi lub obliczalnymi cechami struktury? Szukanie ilościowych zależności między strukturą związków organicznych a reaktywnością ma długą tradycję sięgającą roku 1937, kiedy to pojawiła się publikacja L.P. Hammetta "*The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives*".[15]

W katalizie asymetrycznej problem jest jednak bardziej złożony. O selektywności rzadko decyduje pojedyncza właściwość jednego podstawnika, a raczej jest ona wynikiem jednoczesnego działania efektów sterycznych i elektronowych, przestrzennego dopasowania katalizatora i substratu oraz oddziaływań niekowalencyjnych. Dodatkową trudnością jest to, że nie zawsze wiadomo, którą strukturę układu należy właściwie opisywać. Katalizatory i ich kompleksy z substratem są konformacyjnie labilne, a geometria kompleksu prowadzącego do reakcji może różnić się od najtrwalszej struktury wolnego katalizatora lub substratu.

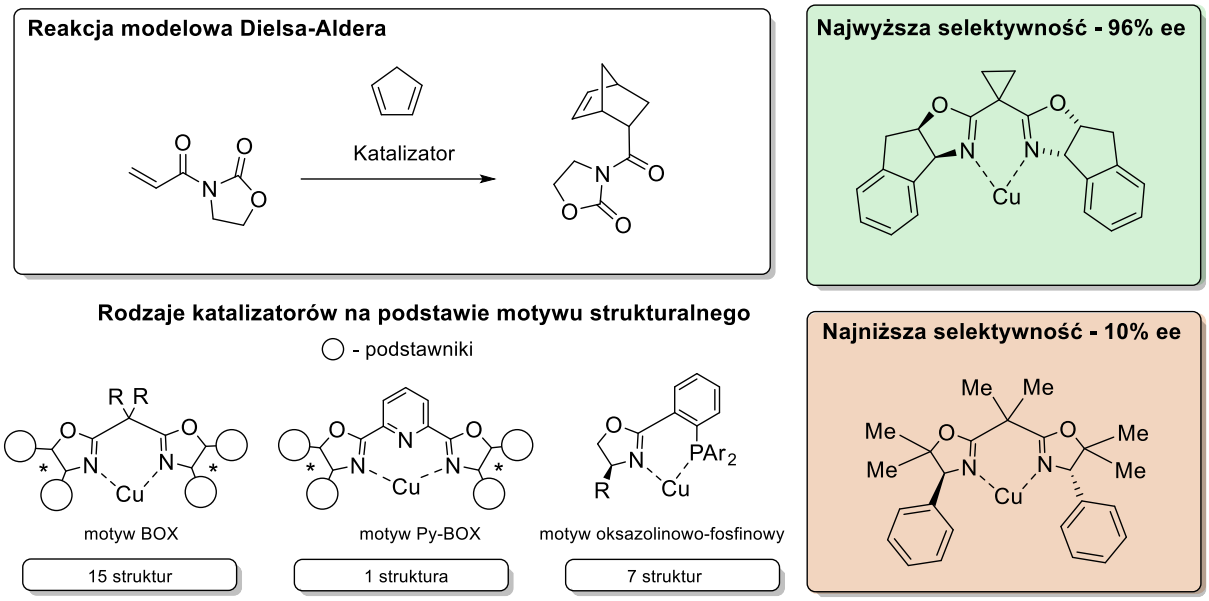
W przypadku katalizy metalami problem pojawia się jeszcze wcześniej, ponieważ nie zawsze oczywista jest sama architektura kompleksu. Przewidywanie sposobu koordynacji liganda do centrum metalicznego pozostaje osobnym zagadnieniem obliczeniowym i dopiero w 2026 roku zaproponowanie strategii pozwalającej przewidywać takie struktury było przedmiotem odrębnej pracy metodologicznej.[43] Zanim więc można opisać liczbowo otoczenie steryczne lub elektronowe katalizatora, trzeba mieć wystarczająco wiarygodny model jego rzeczywistej struktury. Konieczne stało się zatem opracowanie takich reprezentacji numerycznych, które pozwalają uchwycić jednocześnie wiele właściwości cząsteczki i powiązać je z obserwowaną enancjoselektywnością.

4.2.1 Od pól molekularnych do parametryzacji ligandów

Jednym z pierwszych sposobów ilościowego opisu enancjoselektywności było zaadaptowanie do katalizy asymetrycznej metod rozwijanych wcześniej w projektowaniu leków. Punktem wyjścia było założenie, że skoro zmiany struktury katalizatora prowadzą do mierzalnych zmian selektywności reakcji, zależność tę można opisać matematycznie. Takie podejście jest analogiczne do metod QSAR (quantitative structure–activity relationship), w których właściwości strukturalne cząsteczek koreluje się z ich aktywnością biologiczną. W przypadku katalizy asymetrycznej przewidywaną wielkością nie jest jednak aktywność biologiczna, lecz selektywność reakcji, dlatego podejście to określa się również jako QSSR (quantitative structure–selectivity relationship).

Jedną z metod zaczerpniętych z badań nad lekami była CoMFA (comparative molecular field analysis), należąca do metod 3D-QSAR i wymagająca przestrzennego nałożenia analizowanych cząsteczek. W CoMFA cząsteczki umieszcza się w trójwymiarowej siatce, a w jej punktach oblicza się energie oddziaływania ze standardową sondą. Uzyskane w ten sposób pola, przede wszystkim steryczne i elektrostatyczne, są następnie korelowane z badaną wielkością za pomocą regresji częściowych najmniejszych kwadratów (PLS).[44, 45]

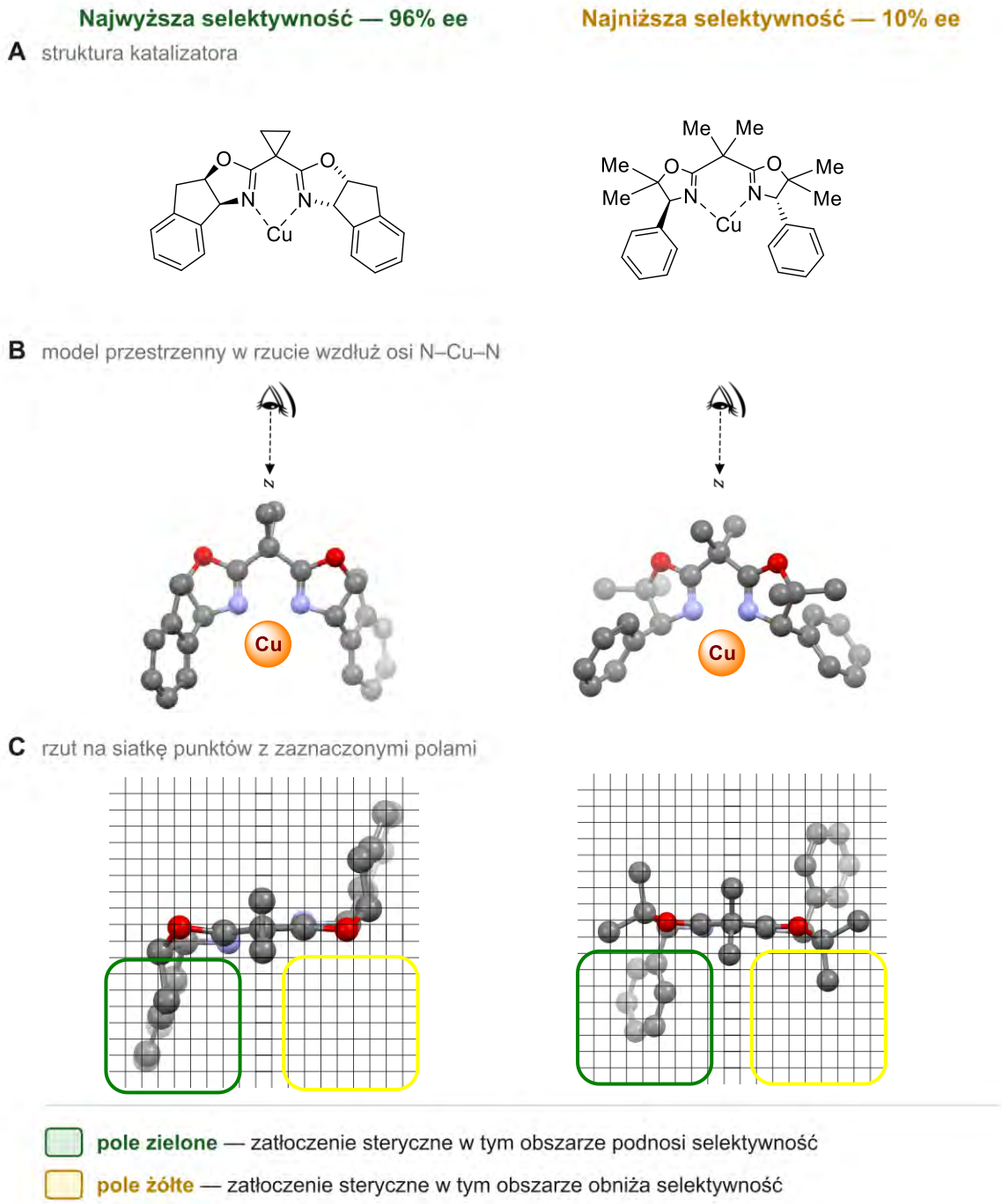
Pierwsze zastosowanie tego podejścia do przewidywania enancjoselektywności katalizatorów opisali Lipkowitz i Pradhan.[46] Modelem była reakcja Dielsa–Aldera *N*-2-alkenoilo-1,3-oksazolidyn-2-onu z cyklopentadienem, katalizowana zestawem 23 kompleksów z ligandami bis(oksazolinowymi) lub fosfinooksazolinowymi (Rysunek 6). Model zwalidowano zarówno wewnętrznie, jak i na zbiorze zewnętrznym, uzyskując w obu przypadkach wysokie wartości q^2. Autorzy pracy skupili się na wyznaczeniu, które cechy strukturalne i elektronowe katalizatora wpływają na podniesienie otrzymywanych wartości nadmiaru enancjomerycznego, a które ją obniżają. Cechy te były bezpośrednio możliwe do przełożenia na projektowanie struktury katalizatora (Rysunek 7). Około 70% zmienności obserwowanego nadmiaru enancjomerycznego przypisano polu sterycznemu, a pozostałą część polu elektrostatycznemu, a więc uzyskano wynik ilościowy odpowiadający intuicji chemicznej, że o indukcji asymetrycznej decyduje przede wszystkim kształt katalizatora. Na tej podstawie sformułowano wskazówki, jak poprawić działanie katalizatorów o niskiej selektywności.



Rysunek 6. *Reakcja modelowa Dielsa–Aldera N-2-alkenoilo-1,3-oksazolidyn-2-onu z cyklopentadienem, katalizowana chiralnymi kompleksami miedzi.*

Znaczenie tej pracy polega na zmianie samego przedmiotu opisu. Zamiast pytać, jak dana cecha podstawnika wpływa na wynik, pyta się, gdzie w przestrzeni wokół katalizatora obecność objętości lub ładunku sprzyja selektywności. Wynikiem jest mapa konturowa, którą chemik może odczytać jako wskazówkę syntetyczną: tutaj dodać objętość, tutaj ją usunąć.

Ograniczenia tego podejścia są jednak wpisane w jego założenia. Wartości pól zależą od wzajemnego ustawienia cząsteczek, co wymusza wyrównanie wszystkich struktur względem wspólnego rdzenia, a więc ogranicza metodę do bibliotek zbudowanych wokół jednego szkieletu. Wymaga ona ponadto wyboru konformeru, a wybór ten jest w praktyce arbitralny.



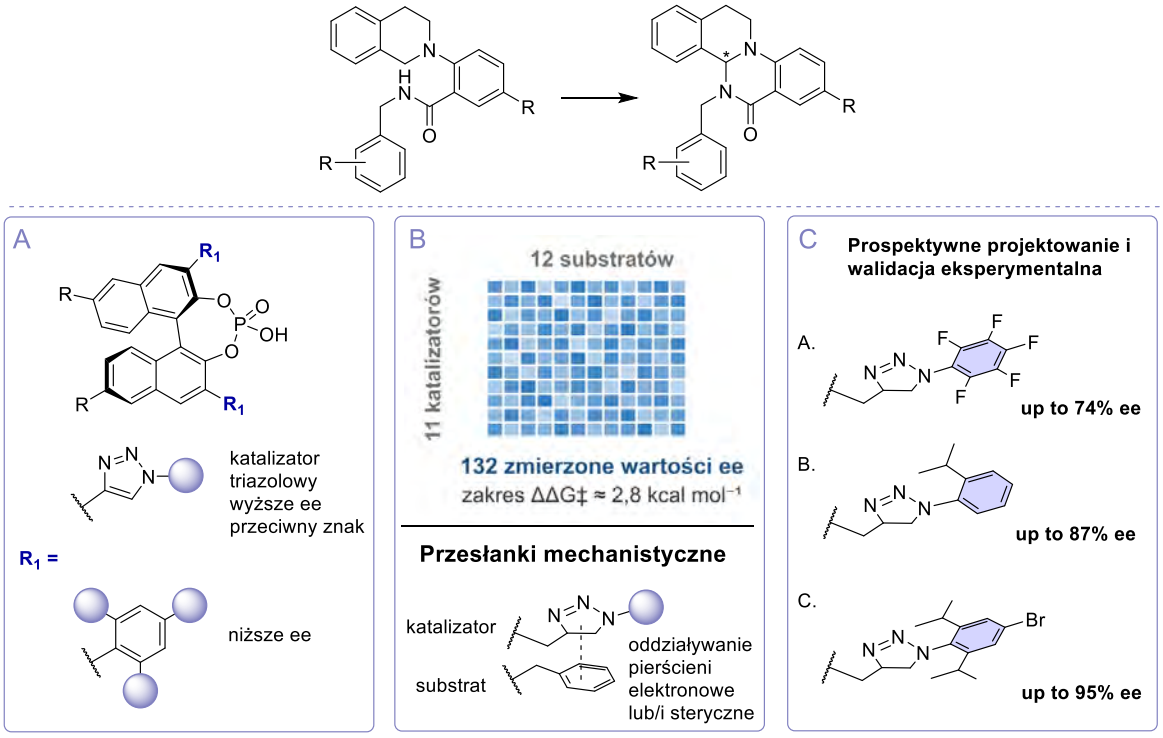
Rysunek 7. *Analiza pól molekularnych zastosowana do katalizatora chiralnego. A. struktury katalizatorów, B. Rzut katalizatora na siatkę, C. Przedstawienie katalizatora w siatce z zaznaczeniem obszarów sprzyjających i niesprzyjających selektywności*

Wariantem tego podejścia, wartym odnotowania ze względu na sposób definiowania obiektu opisu, jest analiza pól molekularnych prowadzona nie dla wolnego katalizatora, lecz dla kompleksu katalizator–substrat odpowiadającego produktowi pośredniemu etapu determinującego enancjoselektywność. Yamaguchi i Sodeoka zastosowali tę strategię do katalizowanego palladem asymetrycznego fluorowania β-ketoestrów.[47] Dla 30 kombinacji pięciu katalizatorów i sześciu substratów pola molekularne obliczano na strukturach Pd-enolanów, gdzie 24 przypadki wykorzystano do budowy modelu, a 6 do jego walidacji. Informacja strukturalna uzyskana z modelu pozwoliła zaprojektować substrat dający 94% *ee*, podczas gdy najwyższa wartość w zbiorze

uczącym wynosiła 81% *ee*. Takie ujęcie lepiej wiąże deskryptor z układem bezpośrednio uczestniczącym w powstawaniu selektywności, ale nie eliminuje problemu konformacyjnego. Wymaga przyjęcia określonego mechanizmu, wskazania odpowiedniego produktu pośredniego oraz optymalizacji obliczeniowej jego struktur. Nie jest to więc powrót do pełnych obliczeń stanów przejściowych, lecz kompromis między opisem czysto korelacyjnym a mechanistycznym.

4.2.2 Model statystyczny jako narzędzie do szukania przesłanek mechanistycznych

Istotnym krokiem w rozwoju modeli korelacyjnych w katalizie asymetrycznej było wykazanie, że mogą one służyć nie tylko do przewidywania enancjoselektywności, lecz także do formułowania i eksperymentalnego testowania hipotez mechanistycznych. Milo, Neel, Toste i Sigman pokazali to w 2015 roku na przykładzie wewnątrzcząsteczkowego dehydrogenacyjnego sprzęgania C–N katalizowanego pochodnymi chiralnych kwasów fosforowych (Rysunek 8).[48] Badany układ stanowił wymagający przypadek, ponieważ niewielkie zmiany struktury katalizatora lub substratu prowadziły do znacznych zmian enancjoselektywności, a jej źródła wiązano z subtelnymi oddziaływaniami niekowalencyjnymi pomiędzy katalizatorem i substratem.



Rysunek 8. *Zastosowanie analizy danych do badania źródeł enancjoselektywności i projektowania katalizatora:* ***(A)*** *systematyczna modyfikacja struktury substratów i chiralnych katalizatorów fosforanowych;* ***(B)*** *parametryzacja ich właściwości strukturalnych i budowa modeli korelacyjnych;* ***(C)*** *prospektywne projektowanie nowych katalizatorów oraz eksperymentalna walidacja przewidywań i hipotez mechanistycznych.*

Autorzy najpierw wykorzystali eksperymenty kinetycznego efektu izotopowego do ustalenia etapu, w którym powstaje enancjoselektywność, a następnie systematycznie zmieniali strukturę obu partnerów reakcji. Zbadano 12 substratów i 11 katalizatorów, otrzymując łącznie 132 wartości enancjoselektywności w zakresie od −54 do 94% *ee*. Strukturę katalizatorów i substratów opisano za pomocą parametrów o znaczeniu fizycznym, m.in. deskryptorów Sterimol, parametrów

geometrycznych oraz obliczonych częstości i intensywności drgań. Następnie za pomocą regresji liniowej identyfikowano parametry skorelowane z ΔΔG‡ i na tej podstawie formułowano hipotezy dotyczące oddziaływań odpowiedzialnych za asymetryczną indukcję. Na podstawie analizy wyniesiono tezę, że jednym z głównych czynników odpowiedzialnych za indukcję asymetryczną jest oddziaływania π-π z udziałem pierścienia triazolowego katalizatora i substratu, którego geometria i siła zależą od właściwości sterycznych i elektronowych obu partnerów. Znaczenie m.in. kąta skręcenia pierścienia triazolowego oraz parametrów drgań pozwoliło przekształcić obserwowane korelacje w konkretne hipotezy dotyczące geometrii kompleksu katalizator–substrat. Hipotezy te następnie sprawdzono prospektywnie, projektując nowe katalizatory i przewidując ich działanie jeszcze przed syntezą (Rysunek 8C). Do tego celu autorzy zbudowali model obejmujący 108 kombinacji dziewięciu katalizatorów arylowych z 12 substratami. Szczególnie istotny był katalizator zaprojektowany tak, aby wymuszać kąt skręcenia pomiędzy pierścieniem triazolowym i arylowym zbliżony do 90°. Zgodnie z przewidywaniami modelu dla wszystkich ośmiu zbadanych substratów uzyskano z jego udziałem najwyższe obserwowane wcześniej wartości enancjoselektywności. Praca ta pokazała zatem możliwość zastosowania modeli korelacyjnych w katalizie asymetrycznej: analiza danych obserwowanej zależności struktura–selektywność przechodzi do stawiania hipotezy mechanistycznej, a następnie do prospektywnego projektowania katalizatora. Kluczowe znaczenie ma przy tym zastosowanie deskryptorów, które można powiązać z rzeczywistymi właściwościami sterycznymi, geometrycznymi lub elektronowymi układu.

4.2.3 Prospektywne kierowanie stereochemią

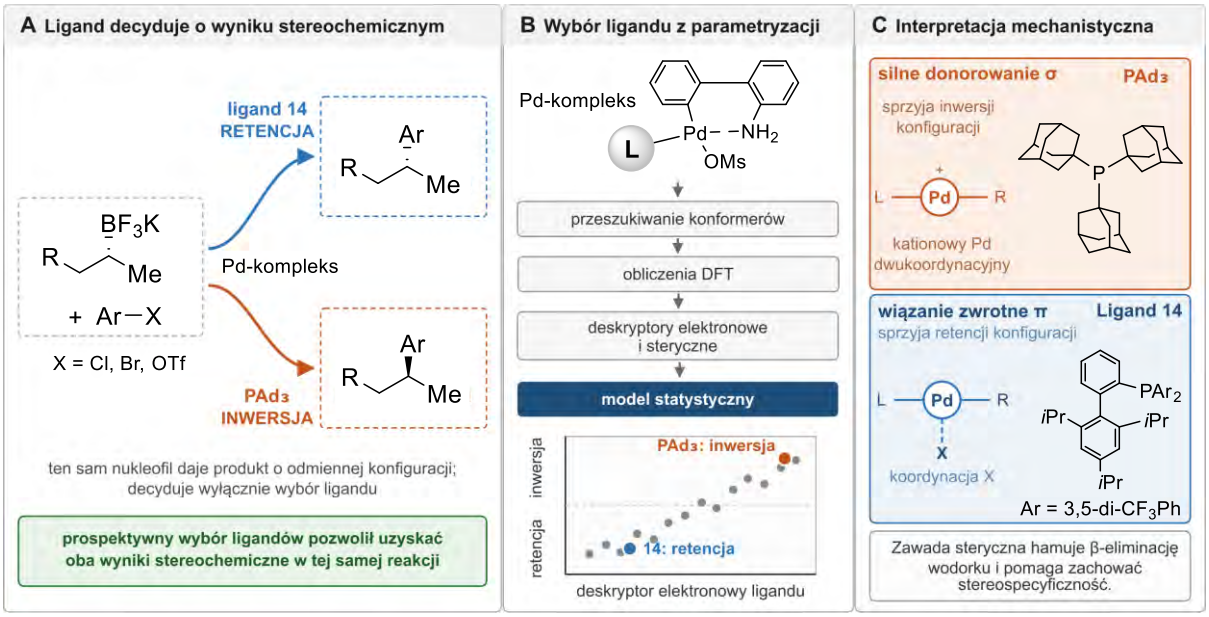
Ważnym krokiem w rozwoju modeli zależności struktura–selektywność było wykorzystanie ich nie tylko do interpretacji istniejących danych, lecz także do projektowania kolejnych eksperymentów. Zhao, Gensch, Biscoe, Sigman i współpracownicy pokazali takie podejście dla stereospecyficznego sprzęgania Suzuki katalizowanego palladem z udziałem enancjomerycznie wzbogaconych drugorzędowych reagentów alkiloboranowych.[49] W tej reakcji etap transmetalacji może przebiegać z retencją albo z inwersją konfiguracji centrum stereogenicznego, prowadząc do przeciwnych wyników stereochemicznych (Rysunek 9A).

Autorzy opisali serię ligandów fosfinowych za pomocą deskryptorów elektronowych i sterycznych wyznaczonych obliczeniowo (Rysunek 9B). Analiza wykazała, że charakter elektronowy fosfiny jest silnie związany z przebiegiem stereochemicznym reakcji: bogate elektronowo trialkilofosfiny sprzyjały ścieżce z inwersją, natomiast mniej elektronodonorowe fosfiny arylowe przesuwały reakcję w stronę retencji. Na podstawie tych zależności wytypowano nowe ligandy, które następnie sprawdzono eksperymentalnie. Odpowiedni dobór fosfiny pozwalał więc otrzymywać produkty o przeciwnej konfiguracji z tego samego enancjomerycznie wzbogaconego reagenta alkiloboranowego (Rysunek 9A).

Dalsza analiza na zbiorze 24 ligandów pozwoliła powiązać obserwowaną stereochemię przede wszystkim z dwoma deskryptorami elektronowymi fosfiny. Autorzy zaproponowali na tej podstawie wyjaśnienie mechanistyczne: silna σ-donacja sprzyjała ścieżce prowadzącej do inwersji, natomiast inne właściwości elektronowe liganda stabilizowały układ prowadzący do retencji (Rysunek 9C). Deskryptory steryczne okazały się natomiast związane przede wszystkim z konkurencyjną β-eliminacją wodorku. Model rozdzielał więc wpływ różnych właściwości liganda na dwa odrębne elementy przebiegu reakcji.

Znaczenie tej pracy polegało przede wszystkim na przejściu od retrospektywnego opisu do prospektywnego wyboru liganda. Model nie odpowiadał już wyłącznie na pytanie, dlaczego dany ligand daje określoną selektywność, lecz pomagał wskazać, jaki ligand należy zastosować, aby

preferowana była retencja albo inwersja konfiguracji. Z punktu widzenia chemika syntetyka oznaczało to możliwość sterowania konfiguracją produktu przez zmianę liganda, bez konieczności zmiany konfiguracji substratu.



Rysunek 9. *Wykorzystanie parametryzacji ligandów do sterowania przebiegiem stereochemicznym sprzęgania Suzuki katalizowanego palladem. (A) Dwie konkurencyjne ścieżki transmetalacji prowadzące do retencji lub inwersji konfiguracji oraz ich związek z właściwościami elektronowymi liganda. (B) Proces modelowania wyboru ligandów (C) Interpretacja wyników modeli i wnioski mechanistyczne wyciągnięte na jej podstawie.*

4.2.4 Podsumowanie

Prace omówione w tym podrozdziale pokazały, że enancjoselektywność można opisać ilościowo jako funkcję obliczalnych cech katalizatora i substratu, a odpowiednio zbudowany model potrafi prowadzić od korelacji do hipotezy mechanistycznej i do prospektywnego projektu. Nie można liczyć jednak na traktowanie tego jako jednoznaczny dowód na dany mechanizm, ale narzędzie wspierające dojście do odpowiedniej hipotezy mechanistycznej. Każde z tych zadań wymagało jednak ponownego dobrania danych, deskryptorów i modelu dla konkretnej reakcji. Naturalnym kolejnym pytaniem stało się zatem, czy informację uchwyconą w jednej reakcji można wykorzystać w innej.

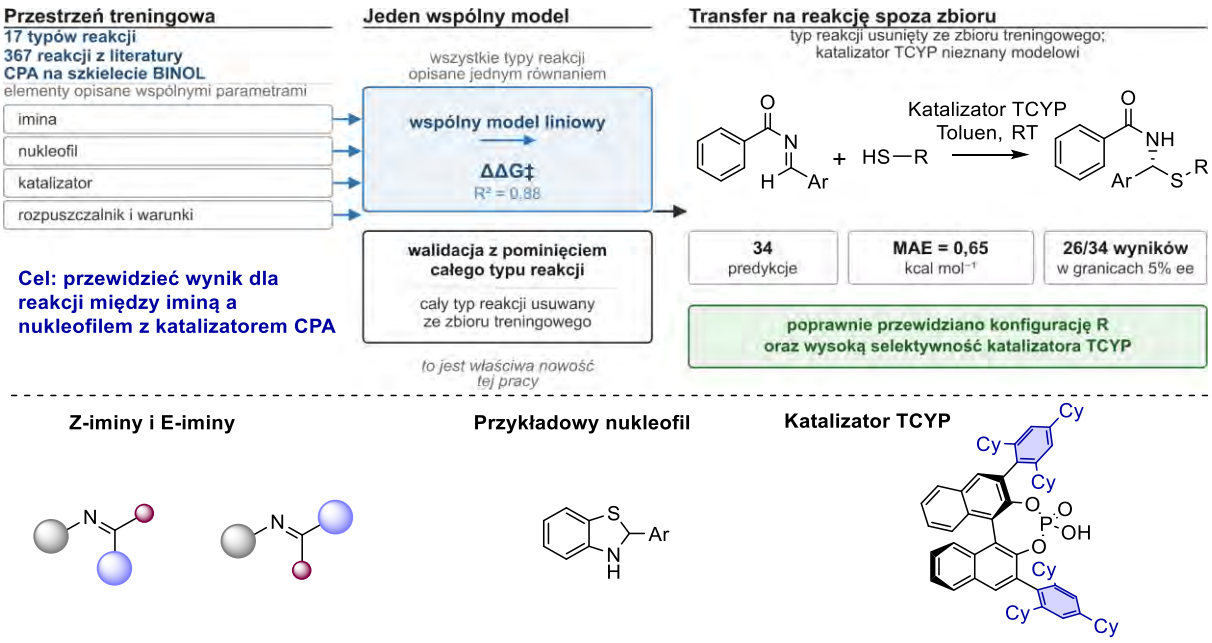
4.3 Transferowalność — czy wiedzę z jednej reakcji można wykorzystać w innej?

Modele opisane wcześniej były projektowane przede wszystkim dla pojedynczej reakcji lub wąskiej rodziny strukturalnej. Dla chemika oznacza to ograniczoną użyteczność: po przejściu do nowego typu substratu lub pokrewnej transformacji nie wiadomo, czy wcześniej wyznaczona zależność nadal obowiązuje. Pytanie tego podrozdziału brzmi więc nie tyle, czy można dokładniej

dopasować kolejny model, lecz czy można znaleźć takie cechy reakcji, które pozostają informatywne również po zmianie części układu chemicznego.

4.3.1 Transferowalność w obrębie reakcji o analogicznym mechanizmie

Punktem wyjścia pracy J. Reid i M. Sigmana[50] była obserwacja znana każdemu chemikowi syntetykowi: dla nowej reakcji wybiera się zwykle warunki, które sprawdziły się w reakcji pokrewnej, i postępowanie to często zawodzi z powodu subtelnych różnic wymagań. Jako układ modelowy wybrano addycje różnych nukleofili do imin katalizowane chiralnymi kwasami fosforowymi pochodnymi BINOL-u, zbierając z literatury 367 reakcji. Istotą tego podejścia nie było więc znalezienie jednego „uniwersalnego parametru katalizatora", lecz zbudowanie wspólnego opisu różnych reakcji, w którym te same rodzaje właściwości fizycznych mogą być porównywane pomiędzy układami (Rysunek 10).



Rysunek 10. *Transfer informacji pomiędzy reakcjami w modelu Reid i Sigmana. Wspólna parametryzacja katalizatora, substratów i warunków pozwoliła zbudować model obejmujący 17 typów reakcji katalizowanych chiralnymi kwasami fosforowymi. Transferowalność sprawdzono przez usuwanie całych reakcji ze zbioru treningowego oraz predykcję danych spoza zbioru, w tym wyników dla katalizatora TCYP i addycji tioli do imin benzoilowych.*

Ważniejszy od samego dopasowania był test leave-one-reaction-out: całą reakcję usuwano ze zbioru treningowego i przewidywano jej wyniki na podstawie pozostałych. Autorzy definiują przy tym pojedynczą reakcję jako zbiór danych pochodzący z jednej publikacji, co czyni ten test operacyjnie podziałem rozłącznym względem źródła literaturowego — jest to jedyny znany mi przykład takiego podziału w katalizie asymetrycznej.

Najmocniejszym sprawdzianem transferowalności była predykcja danych pochodzących z niezależnego badania innego zespołu. Reid i Sigman wykorzystali w tym celu opublikowany wcześniej przez Denmarka i współpracowników zbiór danych dotyczący addycji tioli do imin benzoilowych, w którym wykonano 2150 eksperymentów z użyciem 43 chiralnych kwasów fosforowych oraz macierzy 25 kombinacji nukleofil–elektrofil.[51] Aby sprawdzić rzeczywistą

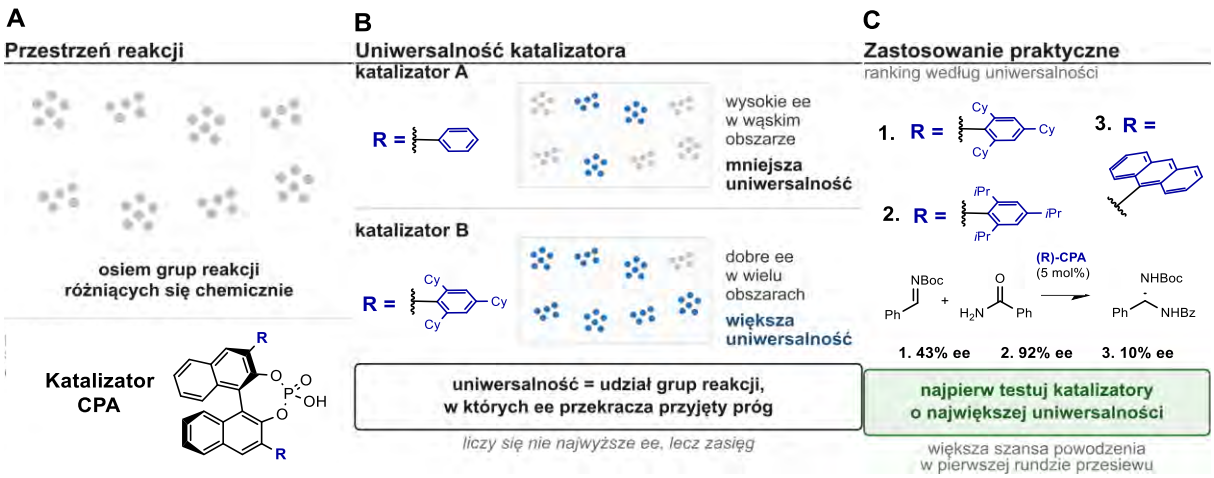
zdolność modelu do przenoszenia informacji między reakcjami, Reid i Sigman usunęli ze swojego zbioru treningowego wszystkie dane dotyczące tej transformacji, a następnie ponownie wytrenowali model. Tak przygotowany model zastosowano do przewidzenia 34 wyników uzyskanych z katalizatorem TCYP (chiralny kwas fosforowy zawierającym grupy 2,4,6-tricykloheksylofenylowe), którego również nie było w zbiorze treningowym. Średni bezwzględny błąd predykcji wyniósł 0,65 kcal mol^{-1}, a 26 z 34 przewidywań różniło się od wartości eksperymentalnych o nie więcej niż 5 punktów procentowych *ee*. Model poprawnie przewidział również konfigurację R produktu oraz wysoką selektywność katalizatora TCYP.

4.3.2 Model pokazujący uniwersalność katalizatora

W praktyce optymalizacji reakcji istotne jest nie tylko to, który katalizator daje najwyższą enancjoselektywność dla pojedynczego substratu, lecz także jak szeroki zakres substratowy toleruje przy zachowaniu dobrej selektywności. Betinol, Lai, Thakur i Reid zaproponowali ilościowy sposób oceny tak rozumianej uniwersalności katalizatorów.[52] Reakcje grupowano według podobieństwa w przestrzeni chemicznej, a następnie oceniano, w ilu chemicznie odrębnych obszarach dany katalizator przekraczał przyjęty próg enancjoselektywności (Rysunek 11A,B).

Metodę zastosowano m.in. do chiralnych kwasów fosforowych. Analiza pokazała, że katalizator często stosowany lub dający wysoką średnią selektywność nie musi być jednocześnie najbardziej uniwersalny. Wśród analizowanych katalizatorów wysoko oceniono TRIP, a model oparty na cechach sterycznych i elektronowych wskazał jeszcze większy potencjał dla TCYP.

Najbardziej praktycznym elementem pracy było wykorzystanie rankingu do wyboru katalizatorów do pierwszego screeningu nowej reakcji (Rysunek 11C). Spośród trzech wysoko ocenionych katalizatorów TRIP dał 92% *ee* już w pierwszej rundzie eksperymentów. Ranking nie wskazywał więc z góry najlepszego katalizatora dla konkretnej reakcji, ale pozwalał racjonalnie wybrać niewielki zestaw struktur, od których warto rozpocząć screening.



Rysunek 11. *Schemat ilościowej oceny uniwersalności katalizatorów i jej wykorzystania w planowaniu screeningu reakcji.* ***(A)*** *Podział przestrzeni reakcji na chemicznie odrębne grupy stanowiące podstawę oceny zakresu stosowalności katalizatora.* ***(B)*** *Porównanie katalizatorów o różnej uniwersalności na przykładzie 3,3'-fenylowego chiralnego kwasu fosforowego i TCYP.* ***(C)*** *Wykorzystanie rankingu uniwersalności do wyboru katalizatorów do eksperymentalnego screeningu addycji benzamidu do arylowej N-Boc iminy;1. TCYP,2. TRIP i 3. pochodna 9-fenantrylowa dały odpowiednio 43, 92 i 10% ee.*

4.3.3 Tworzenie modeli w oparciu o reprezentacje struktur przejściowych decydujących w etapie stereoindukcji

W katalizie asymetrycznej dokładność modelu zależy nie tylko od zastosowanego algorytmu, lecz również od tego, jak zakodowana zostanie informacja chemiczna. Gallarati i współpracownicy pokazali to dla propargilacji benzaldehydu katalizowanej chiralnymi *N,N′*-di-tlenkami bipirydyny.[53] Zamiast opisywać pojedynczy związek, porównali deskryptory SLATM dwóch intermediatów znajdujących się bezpośrednio przed i po stanie przejściowym determinującym enancjoselektywność. Reprezentacja SLATM, utworzona przez odjęcie obu wektorów, zachowywała przede wszystkim informację o zmianach strukturalnych zachodzących podczas tego etapu reakcji. Pozwoliło to zmniejszyć błąd przewidywania energii aktywacji z 0,54 kcal mol^{-1} dla reprezentacji pojedynczego intermediatu do 0,31 kcal mol^{-1}, a po selekcji cech do około 0,25 kcal mol^{-1}. Wynik ten pokazał, że w przewidywaniu stereoselektywności korzystniejsze może być opisanie zmiany zachodzącej w reakcji, a nie tylko struktur jej poszczególnych elementów. Podejście wymaga jednak znajomości mechanizmu i wskazania odpowiednich struktur przejściowych etapu enancjoróżnicującego.

Inny problem rozwiązali Schoepfer i współpracownicy, opracowując jednolity sposób opisu chiralnych ligandów dwukleszczowych, który nie wymaga definiowania nowego zestawu deskryptorów dla każdej reakcji.[54] Każdy ligand opisano za pomocą 232 cech elektronowych, sterycznych i topologicznych, a następnie dla każdej z czterech reakcji automatycznie wybierano maksymalnie trzy cechy i budowano model bayesowskiej regresji grzbietowej. Co istotne, modele powstały na bardzo małych zbiorach obejmujących zaledwie 19–30 wyników screeningu ligandów na reakcję, przy stałych reagentach i zmienianym przede wszystkim ligandzie. Dla wszystkich czterech reakcji błąd walidacji leave-one-out nie przekraczał 0,29 kcal mol^{-1}, a w trzech z czterech modeli szczególnie istotne okazały się cechy topologiczne związane m.in. ze sztywnością liganda.

Bayesowska regresja dostarczała ponadto oszacowania niepewności, dzięki czemu model można było wykorzystać do wyboru kolejnych ligandów w optymalizacji bayesowskiej. W retrospektywnej analizie reakcji oksyalkinylacji najlepszy ligand został w ten sposób odnaleziony po 6 zamiast 19 eksperymentów. Co byłoby imponujące tylko w przypadku analizy prospektywnej. Prace te pokazują zatem dwa różne sposoby radzenia sobie z ograniczoną ilością danych: reprezentację opartą na zmianie zachodzącej podczas konkretnego etapu reakcji oraz ogólny opis jednego zmiennego elementu układu katalitycznego. Jednocześnie zakres obu podejść pozostaje ściśle określony przez rodzaj informacji dostępnej w danym zbiorze.

W pracy Reid i Sigmana[50] omówionej w podrozdziale 4.3.1 wspólną reprezentację różnych reakcji tworzono przez oddzielny opis substratów, katalizatora i warunków, a następnie połączenie ich deskryptorów w jeden wektor (Rysunek 12A). Pozwalało to przenosić informacje pomiędzy różnymi kombinacjami reagentów i katalizatorów, ale nie opisywało bezpośrednio oddziaływań powstających pomiędzy nimi w trakcie reakcji.

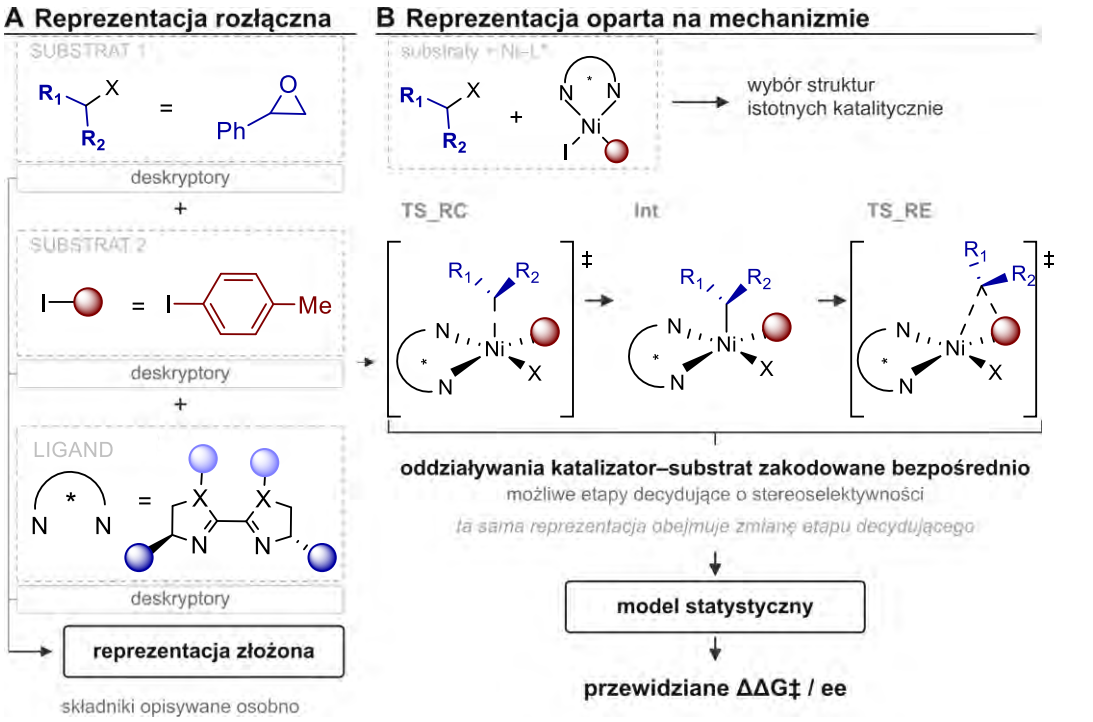
Tym razem Sigman oraz Gallarati, Bucci i Doyle zaproponowali inne podejście dla enancjoselektywnych sprzęgań C(sp^3) katalizowanych niklem.[55] Zamiast opisywać wolny ligand i substraty osobno, wyprowadzili deskryptory ze struktur odpowiadających kolejnym etapom cyklu katalitycznego: stanowi przejściowemu wychwytu rodnika (TSRC), intermediatowi (Int) oraz stanowi przejściowemu eliminacji redukcyjnej (TSRE) (Rysunek 12B). Taka reprezentacja zawiera więc jednocześnie informację o katalizatorze, substratach i ich wzajemnym ułożeniu w strukturach istotnych dla przebiegu reakcji.

Podejście to łączy wiedzę mechanistyczną z modelowaniem statystycznym. Nie oblicza się bezpośrednio różnicy energii stanów przejściowych prowadzących do obu enancjomerów. Zamiast

tego z zespołów konformerów TSRC, Int i TSRE wyznacza się deskryptory steryczne, geometryczne i elektronowe, a model uczy się ich związku z eksperymentalną enancjoselektywnością. Co istotne, najlepsze wyniki dla badanego układu dawały deskryptory TSRE — etapu wcześniej wskazanego w analizie mechanistycznej jako stereodeterminujący. Porównanie reprezentacji pochodzących z różnych etapów cyklu może więc dostarczać dodatkowych przesłanek dotyczących miejsca, w którym powstaje kontrola stereochemiczna.

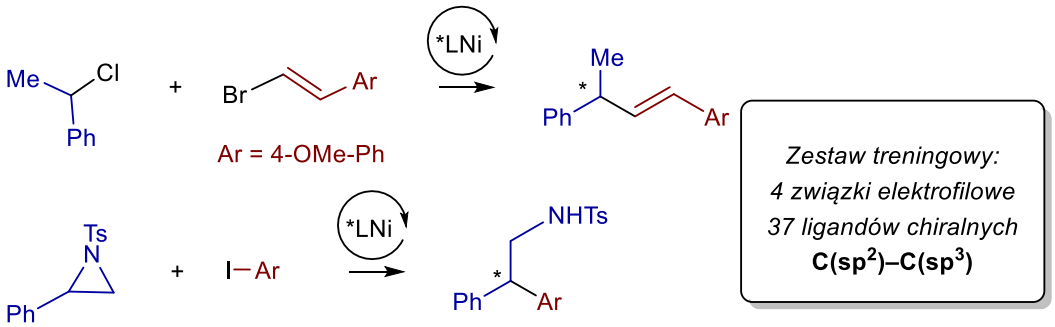
Najważniejszym elementem pracy było jednak sprawdzenie, czy taki sposób opisu pozwala przenosić informacje pomiędzy różnymi, ale mechanistycznie pokrewnymi reakcjami. Model zbudowany dla kilku sprzęgań niklowych $C(sp^2)$–$C(sp^3)$ zastosowano następnie do sprzęgania $C(sp^3)$–$C(sp^3)$, należącego do innego typu reakcji i obejmującego inną przestrzeń ligandów (Rysunek 12C). Pomimo tego uzyskano MAE = 0,18 kcal mol^{-1} i RMSE = 0,24 kcal mol^{-1}.[55]

W porównaniu z podejściem we współpracy z Reid (podrozdział 4.3.1.) zmienia się więc przede wszystkim wspólny „język" opisu reakcji. Tam były nim deskryptory oddzielnych składników układu (Rysunek 12A), tutaj są to deskryptory analogicznych struktur występujących w przebiegu cyklu katalitycznego (Rysunek 12B). Dzięki temu informacje z małych zbiorów kilku reakcji można było wykorzystać do przewidywania innej, mechanistycznie pokrewnej transformacji (Rysunek 12C). Ceną tej transferowalności jest konieczność wcześniejszego sformułowania wiarygodnej hipotezy mechanistycznej i wygenerowania odpowiednich struktur pośrednich i stanów przejściowych.

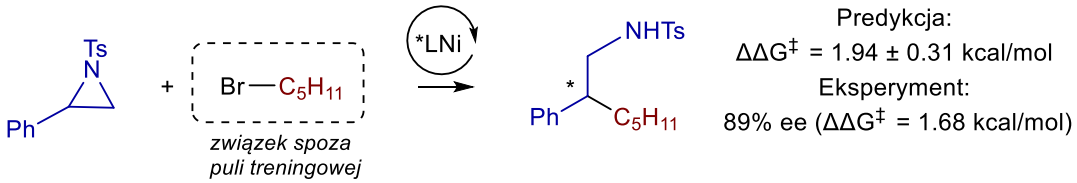


C Transfer na nową reakcję

Model uczy się na TS_RC, Int, TS_RE sprzęgania C(sp^2)–C(sp^3)



Zewnętrzny test - reakcja sprzęgania C(sp^3)–C(sp^3)



Model wybrał ligand spoza wcześniej wybieranych do sprzęgania C(sp^2)–C(sp^3)

Rysunek 12. *Mechanistycznie informowana reprezentacja reakcji umożliwiająca transfer modeli enancjoselektywności z małych zbiorów danych. (A) W podejściu holistycznym składniki reakcji są parametryzowane oddzielnie, a ich deskryptory łączone we wspólną reprezentację. (B) W podejściu Gallarati i wsp. deskryptory wyprowadzane są z katalitycznie istotnych struktur TSRC, Int i TSRE, dzięki czemu reprezentacja bezpośrednio uwzględnia oddziaływania katalizator–substrat oraz możliwość zmiany etapu odpowiedzialnego za stereoindukcję. (C) Przykład transferu między reakcjami: model uczony na sprzęganiach C(sp²)–C(sp³) obejmujących cztery typy elektrofilów i 37 chiralnych ligandów zastosowano do zewnętrznego testu na niewidzianej reakcji H, będącej sprzęganiem C(sp³)–C(sp³).*

4.3.4 Podsumowanie

Ponieważ każda z metod wymaga hipotezy mechanistycznej na wejściu, przed modelowaniem trzeba zadecydować, jakie stany przejściowe wygenerować. Metody te wymagają również znacznie większej liczby obliczeń strukturalnych niż parametryzacja wolnego liganda, substratu lub produktu.

Kiedy widzimy jak możemy opracować nowe reakcje korzystając z analogii dostępnych rekordów, naturalnym kolejnym pytaniem jest czy model może wskazać strukturę katalizatora, której nie ma jeszcze w dostępnej bibliotece. To prowadzi bezpośrednio do problemu projektowania nowych katalizatorów.

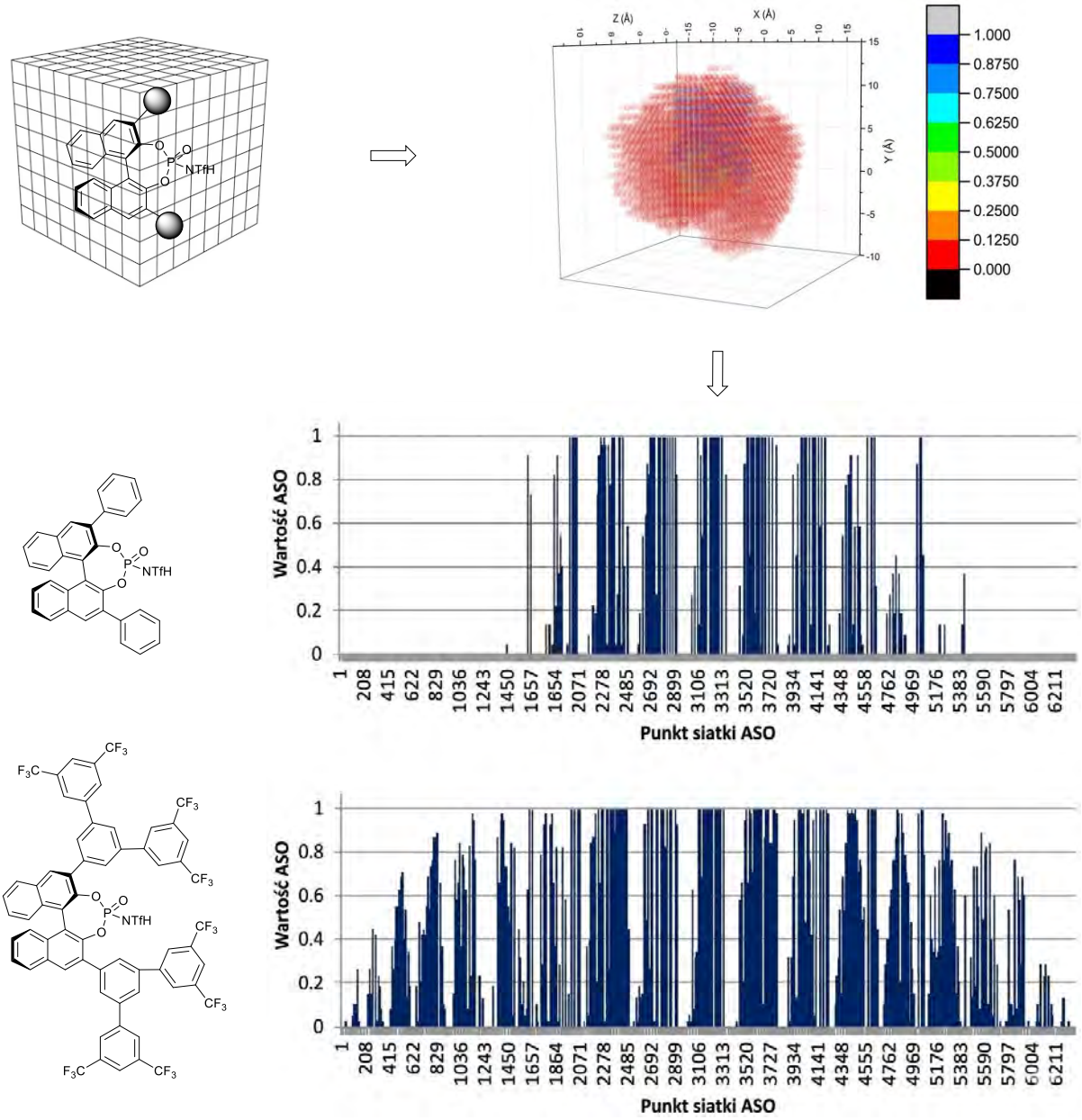
4.4 Projektowanie katalizatorów — czy model może wskazać nową strukturę katalizatora?

Ograniczenie opisane w poprzednim podrozdziale skłania do postawienia pytania odwrotnego. Zamiast przenosić model między reakcjami, można pozostać przy jednej reakcji i jednej rodzinie katalizatorów, ale zażądać od modelu czegoś więcej: wskazania modyfikacji strukturalnej katalizatora, która podniesie jego efektywność. Zadanie to jest trudniejsze pod jednym istotnym względem. Model korelacyjny wystarczy oceniać na tym, jak dobrze odtwarza dane, a model projektujący musi zostać sprawdzony syntezą. Jego predykcja albo się potwierdzi, albo nie.

4.4.1 Pierwszy prospektywnie potwierdzony projekt

Przełomowym przykładem tego podejścia jest praca zespołu Denmarka dotycząca addycji tioli do *N*-acyloimin katalizowanej chiralnymi kwasami fosforowymi.[51] Reakcja była wówczas dobrze scharakteryzowana, a mimo to wybór katalizatora sprowadzał się do przeszukania dostępnej biblioteki i przeprowadzeniu eksperymentów testowych. Autorzy pracy postawili sobie za cel opracować model, który zaproponuje jeszcze bardziej selektywny katalizator niż te, które są obecnie dostępne.

Zbiór danych zbudowano w całości *in silico*. Z 403 syntetycznie dostępnych fragmentów wygenerowano bibliotekę 806 chiralnych kwasów fosforowych, opisując każdą strukturę deskryptorami liczonymi na siatce punktów otaczających cząsteczkę, kluczowym z nich była średnie zatłoczenie steryczne (ASO, ang. average steric occupancy), uzupełnione o deskryptory elektronowe oraz pola oddziaływań (Rysunek 13). Reprezentacja ta uwzględnia rozkład konformerów, a nie pojedynczą strukturę, co odróżnia ją od analizy CoMFA omówionej w podrozdziale 4.2.1. Podzbiór treningowy, złożony ze struktur CPA, wybrano algorytmicznie tak, aby możliwie równomiernie pokrywał przestrzeń deskryptorową i nazwano go uniwersalnym zbiorem treningowym (UTS). Został zsyntezowany jeden raz przed poznaniem jakichkolwiek wyników eksperymentalnych, a jego uniwersalność polega na tym, że może posłużyć do optymalizacji dowolnej reakcji z katalizatorem opartym na motywie strukturalnym chiralnego kwasu fosforowego (CPA).



Rysunek 13. *Przełożenie struktury przestrzennej katalizatora na wartości deskryptora średniego zatłoczenia sterycznego (ASO).*

Model oceniono na trzech niezależnych zbiorach zewnętrznych o rosnącej trudności, gdzie autorzy weryfikowali działanie modelu dla nowych kombinacji substratów, nowych katalizatorów (dla znanych kombinacji substratów) oraz dla nowych kombinacji substratów i nowych katalizatorów. Dla nowych produktów z katalizatorami ze zbioru treningowego średnie odchylenie bezwzględne wyniosło 0,161 kcal mol^{-1}, dla katalizatorów spoza zbioru treningowego przy znanych kombinacjach substratów — 0,211 kcal mol^{-1}, a dla nowych produktów i nowych katalizatorów jednocześnie — 0,236 kcal mol^{-1}. Najmocniejszy jest jednak drugi eksperyment: po usunięciu ze zbioru treningowego wszystkich reakcji o selektywności powyżej 80% *ee* głęboka sieć neuronowa mimo to odtworzyła dane eksperymentalne i poprawnie wskazała katalizatory najbardziej selektywne, a więc obszar, którego w danych treningowych nie było. Wynikiem końcowym było przewidzenie i eksperymentalne potwierdzenie katalizatorów o wyższej selektywności niż jakikolwiek katalizator obecny w zbiorze treningowym.

Nowość tego podejścia polega na połączeniu trzech elementów: wirtualnej biblioteki struktur, celowo dobranego zbioru treningowego oraz prospektywnej walidacji. Dzięki temu model nie tylko opisuje znane wyniki, ale może wskazywać nowe struktury warte sprawdzenia eksperymentalnego. Bariera wstępu w takie badania jest jednak znaczna: koszt wygenerowania i zoptymalizowania zbiorów konformerów dla 806 struktur, ich wzajemne wyrównanie przestrzenne oraz obliczenie deskryptorów siatkowych trzeba ponieść od nowa dla każdej nowej klasy katalizatorów i każdej nowej reakcji, a samo wyrównanie wymaga wspólnego motywu strukturalnego katalizatora, co zawęża metodę do bibliotek zbudowanych wokół jednego szkieletu. Sądzę, że praca ta jest cytowana w znacznej mierze za wytworzony zasób i za wzorzec postępowania, a nie za samą metodę modelowania, która pozostaje regresją na deskryptorach.

Ponieważ głównym ograniczeniem pozostaje koszt obliczenia deskryptorów dla dużych bibliotek ligandów, największe zapotrzebowanie jest na powstawanie bibliotek z umieszczeniem gotowych, obliczonych z wyprzedzeniem reprezentacji dla całej klasy związków. Z takim motywem przewodnim powstała platforma Kraken, w której zespoły Sigmana i Aspuru-Guzika, na podstawie dokładnych obliczeń kwantowo-mechanicznych dla 1558 ligandów, zbudowali model i zrobili predykcje właściwości dla około 300 000 nowych ligandów fosforoorganicznych. Każdy z nich został opisany za pomocą ponad 1500 deskryptorów uwzględniających również różnice pomiędzy konformerami.[32] Praca ta nie dotyczy bezpośrednio katalizy asymetrycznej: nie zbudowano w niej modelu selektywności ani nie przeprowadzono walidacji eksperymentalnej. Jej znaczenie polega na stworzeniu gotowego zasobu obliczeniowego, który można wykorzystać w późniejszych badaniach nad kompleksami zawierającymi takie ligandy, bez konieczności każdorazowego generowania całego zestawu deskryptorów od podstaw.

Warto wspomnieć o takich zasobach również z szerszej perspektywy rozwoju tej dziedziny. Budowa wysokiej jakości zbiorów danych i ich parametryzacja należą do najbardziej czasochłonnych i kosztownych etapów pracy w chemii opartej na danych. Obejmują nie tylko zebranie wyników eksperymentalnych, ale także ich standaryzację, kontrolę jakości, przygotowanie struktur oraz obliczenie reprezentacji molekularnych. Dlatego tworzenie możliwie dużych, dobrze opisanych i wielokrotnego użytku baz danych może istotnie przyspieszać rozwój metod predykcyjnych: kolejne zespoły mogą koncentrować się na budowie i porównywaniu modeli, zamiast za każdym razem odtwarzać ten sam etap przygotowania danych. W tym sensie zasoby takie jak Kraken mają znaczenie jako infrastruktura umożliwiająca szybszy i bardziej porównywalny rozwój badań nad projektowaniem katalizatorów.

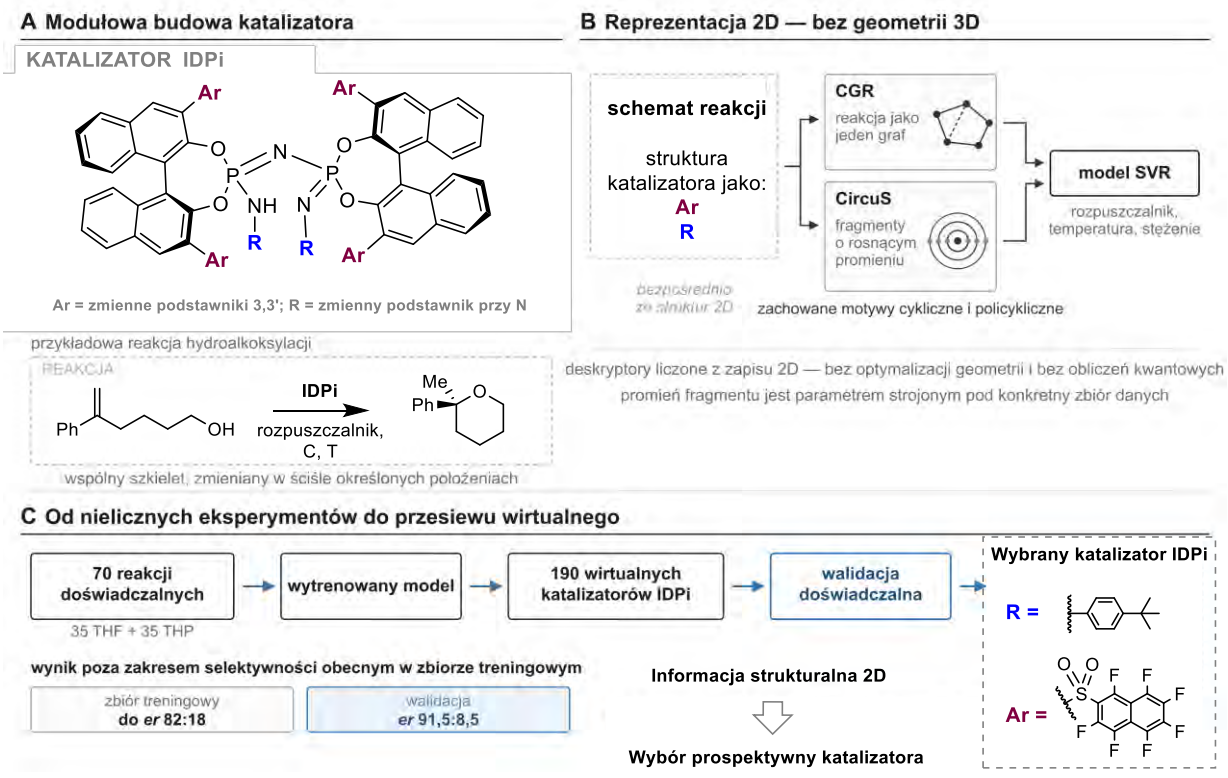
4.4.2 Projektowanie katalizatora w oszczędnym wydaniu

Inną strategią ograniczenia kosztu parametryzacji jest odejście od deskryptorów wymagających generowania i analizy geometrii trójwymiarowych. Tsuji, Sidorov, Varnek, List i współpracownicy zastosowali deskryptory fragmentowe wyprowadzane bezpośrednio ze struktur dwuwymiarowych do przewidywania enancjoselektywności reakcji katalizowanych katalizatorami typu IDPi (Rysunek 14A), których strukturę modyfikuje się przez zmianę podstawników Ar w pozycjach 3,3′ jednostek BINOL oraz podstawnika R przy atomie azotu (Rysunek 14A).[14]

Początkowy zbiór obejmował 35 reakcji hydroalkoksylacji prostych alkenów prowadzących do 2,2-dipodstawionych tetrahydrofuranów (THF), dla których stosunek enancjomerów mieścił się w zakresie od 15:85 do 91:9. Dane nie obejmowały więc katalizatorów o bardzo wysokiej selektywności. Stanowiło to wymagające zadanie prospektywne, ponieważ model miał wskazać katalizatory osiągające selektywność wyższą niż wartości reprezentowane w zbiorze treningowym.

Substraty i produkty kodowano w postaci skondensowanego grafu reakcji (CGR), który łączy je w jeden graf zawierający informację o wiązaniach zmieniających się podczas reakcji (Rysunek 14B). Katalizator mógł być reprezentowany jako cała struktura lub jedynie przez zmienne podstawniki Ar i R, co ograniczało udział niezmiennego rdzenia w reprezentacji. Szczególne znaczenie miało opracowanie deskryptorów CircuS (ang. circular substructures): klasyczne fingerprinty o zadanym promieniu (zasięgu) opisują lokalne otoczenie atomu, lecz nie zachowują wprost informacji o całych zamkniętych pierścieniach, co ogranicza ich zdolność do rozróżniania rozbudowanych układów aromatycznych i podstawników policyklicznych, częstych w strukturach katalizatorów. Do reprezentacji włączono także parametry fizykochemiczne rozpuszczalnika, a wpływ temperatury i stężenia uwzględniono, przeliczając dane do wspólnych warunków przed modelowaniem. Regresję prowadzono metodą wektorów nośnych (SVR), optymalizując przestrzeń deskryptorów i hiperparametry algorytmami genetycznymi z dziesięciokrotną walidacją krzyżową.

Najlepszy model oparty na deskryptorach CircuS osiągnął dla zbioru wyjściowego MAE 0,741 kJ mol^{-1}, a dla wspólnego zbioru 35 przykładów tetrahydrofuranowych i 35 tetrahydropiranowych - 0,754 kJ mol^{-1}. Na tej podstawie przeprowadzono wirtualne przeszukanie 190 katalizatorów odpowiadających kombinacjom podstawników nieobecnym w danych eksperymentalnych (Rysunek 14C). Najbardziej selektywny z przetestowanych katalizatorów dał stosunek enancjomerów 91,5:8,5, podczas gdy najwyższa selektywność w zbiorze treningowym wynosiła 82:18.



Rysunek 14. *Schemat zastosowania deskryptorów 2D do projektowania katalizatorów IDPi: (A) modułowa struktura katalizatora i zmienne podstawniki Ar i R; (B) reprezentacja reakcji za pomocą CGR i deskryptorów CircuS oraz budowa modelu SVR; (C) wirtualny screening nowych kombinacji podstawników i ich eksperymentalna walidacja.*

Interpretacja takich wyników wymaga sprawdzenia, czy model rzeczywiście wykorzystuje informację chemiczną zawartą w deskryptorach. Znaczenie takich kontroli pokazał komentarz Chuanga i Keisera do pracy Ahnemana i współpracowników o przewidywaniu wydajności sprzęgania Buchwalda–Hartwiga [56] wspomniany w rozdziale 3.4.3.[40] W pracy Tsuji i

współpracowników przeprowadzono analogiczne modele kontrolne, gdzie reprezentację CircuS porównano z kodowaniem one-hot podstawników oraz z cechami losowymi, i oba warianty dawały wyraźnie słabsze wyniki, co wspiera interpretację, że model nie opierał się wyłącznie na rozpoznawaniu kombinacji podstawników obecnych w danych.

Zakres stosowalności reprezentacji sprawdzono dodatkowo na niezależnym zbiorze z pracy Denmarka opisanej w podrozdziale 4.4.1. W tym teście model oparty na CircuS, szczególnie dla zewnętrznego zbioru katalizatorów, dawał wyniki bliższe rezultatowi raportowanemu dla deskryptorów trójwymiarowych niż pozostałe porównywane reprezentacje dwuwymiarowe. Nie oznacza to, że reprezentacje dwuwymiarowe mogą ogólnie zastąpić deskryptory 3D, pokazuje jednak, że dla części problemów możliwe jest dobrze dobrać deskryptory fragmentowe, które stanowią znacznie tańszą alternatywę, możliwą do obliczenia z notacji SMILES oraz bez generowania konformerów i obliczeń kwantowochemicznych.

4.5 Wybór katalizatora z wcześniej zdefiniowanej przestrzeni

4.5.1 Dopasowanie warunków reakcji do cech substratu

„Mam nową parę reagentów i skończony zestaw ligandów/warunków. Od czego zacząć?”

Dobrym przykładem wyboru katalizatora spośród wcześniej zdefiniowanych możliwości jest narzędzie opracowane przez Rineharta, Denmarka i współpracowników dla sprzęgania Buchwalda–Hartwiga.[57] Reakcja ta nie jest asymetryczna, a przewidywanym wynikiem jest wydajność, jednak problem wyboru ma podobną strukturę: dla nowej pary substratów chemik musi wybrać odpowiedni ligand oraz warunki reakcji. Autorzy ograniczyli przestrzeń do 20 znanych ligandów fosfinowych, trzech zasad i trzech rozpuszczalników, co dawało 180 możliwych zestawów warunków dla każdej pary reagentów. Zadaniem modelu nie było więc zaprojektowanie nowego katalizatora, lecz wskazanie, które spośród dostępnych kombinacji warto sprawdzić eksperymentalnie.

Kluczowym elementem pracy było przygotowanie danych specjalnie do tego zadania. Autorzy wskazali, że istniejące dane literaturowe i patentowe obejmują wiele różnych reakcji, ale nie tworzą wystarczająco systematycznej mapy zależności między strukturą obu reagentów a wyborem liganda i warunków. Zdefiniowana przez nich przestrzeń 50 nukleofili, 50 bromków, 20 ligandów, trzech zasad i trzech rozpuszczalników obejmowała 450 000 możliwych eksperymentów, dlatego nie mogła zostać przebadana doświadczalnie. Zamiast tego kolejne pary substratów i warunki wybierano iteracyjnie, wykorzystując modele oraz wiedzę eksperta. W efekcie przebadano około 3300 reakcji, obejmujących 121 z 2500 możliwych par substratów, a uzyskane dane służyły do stopniowego rozszerzania obszaru, w którym model mógł dokonywać predykcji.

Praktyczny sens narzędzia był prosty: po podaniu nowej pary reagentów model przewidywał wyniki dla 180 dostępnych kombinacji warunków i wskazywał najbardziej obiecujące z nich jako punkt startowy do eksperymentu. Nie miał zastępować późniejszej optymalizacji, lecz ograniczyć początkowe poszukiwanie warunków metodą prób i błędów. W walidacji obejmującej 13 nowych par reagentów eksperymentalnie sprawdzano jedynie niewielką liczbę najwyżej ocenionych warunków. Dla 10 przypadków najwyższa uzyskana wydajność przekraczała 85%. Autorzy pokazali również przykłady, w których model zawodził dla substratów wprowadzających rodzaj reaktywności nieobecny wcześniej w danych. Praca dobrze ilustruje więc zarówno możliwości, jak i ograniczenie strategii wyboru spośród znanych opcji: model może skutecznie uporządkować

dostępne ligandy i warunki, ale jego użyteczność zależy od tego, czy odpowiednia chemia jest reprezentowana w zbiorze treningowym.

„Mam ogromną, ale wcześniej zdefiniowaną przestrzeń możliwych ligandów. Które warto w ogóle syntetyzować?”

W obrębie katalizy asymetrycznej podobne podejście zastosowali Olen, Zahrt, Denmark i współpracownicy do wyboru liganda w winylogowej reakcji aldolowej Mukaiyamy katalizowanej kompleksami miedzi z bis(oksazolinami) (BOX).[58] Punktem wyjścia była zdefiniowana *in silico* biblioteka 96 110 ligandów BOX, obejmująca struktury handlowo dostępne i możliwe do otrzymania znanymi drogami syntetycznymi. Każdy ligand opisano zależnymi od konformacji deskryptorami sterycznymi, a następnie z tej przestrzeni wybrano 26 możliwie zróżnicowanych struktur, które zsyntezowano i przebadano dla pięciu aldehydów. Jeden z ligandów tej początkowej biblioteki osiągnął średnio 88% *ee*, przewyższając ligandy znalezione we wcześniejszym screeningu komercyjnych katalizatorów.

Na podstawie uzyskanych danych porównano następnie trzy sposoby wskazywania kolejnych kandydatów: ilościowe modelowanie zależności struktura–selektywność, wybór najbliższych sąsiadów najbardziej selektywnego liganda oraz klastrowanie jego analogów. Łącznie wybrano i zsyntezowano kolejnych 12 ligandów. W tej pracy lepiej sprawdziło się szukanie nowych ligandów przez podobieństwo strukturalne do sprawdzonego, skutecznego liganda niż próba przewidywania *ee* z modelu QSSR. Takie lokalne przeszukiwanie pozwoliło poprawić wynik dla czterech z pięciu produktów, osiągając do 98% *ee*, podczas gdy model ilościowy słabo przewidywał selektywność całkowicie nowych ligandów. Metody oparte na statystyce i podobieństwie okazały się zarówno skuteczniejsze, jak i łatwiejsze do zastosowania niż pełne modelowanie QSSR, pozwalając poprawić najlepszy wynik dla czterech z pięciu produktów, do 98% *ee*.

„Mam bibliotekę ligandów i kilka kryteriów jednocześnie — aktywność, regio- i stereoselektywność. Który ligand może dać oczekiwany rezultat?”

Innym przykładem wyboru liganda z wcześniej zdefiniowanej przestrzeni jest praca zespołu Sigmana we współpracy z Genentech i Roche dotycząca dwóch kolejnych etapów procesu syntezy substancji aktywnej z zastosowaniem w leczeniu astmy. [59] Dwie optymalizowane reakcje były katalizowane metalem szlachetnym w połączeniu z chiralnym ligandem bisfosfinowym - były to katalizowana palladem reakcja Hayashiego-Hecka oraz katalizowana rodem reakcja hydroformylowania alkenów. Problem był wielowymiarowy: ligand powinien zapewniać odpowiednią reaktywność, a jednocześnie wysoką regio- i stereoselektywność. Autorzy przygotowali bibliotekę ponad 550 ligandów bisfosfinowych, obejmującą związki handlowo dostępne, znane z literatury oraz dostępne syntetycznie, i opisali je deskryptorami wyznaczonymi na podstawie obliczeń DFT. Dla obu reakcji zastosowano dwuetapową strategię: najpierw model klasyfikacyjny określał, które ligandy mają szansę utworzyć aktywny katalizator, a następnie tylko dla tej części przestrzeni modelowano selektywność za pomocą regresji liniowej. Wirtualny screening całej biblioteki wskazał następnie ligandy o korzystnym połączeniu obu właściwości. Ich eksperymentalna walidacja poprawiła przede wszystkim regioselektywność obu etapów przy zachowaniu wysokiej reaktywności i stereoselektywności, a wybrane warunki potwierdzono również w skali gramowej.

Istotnym elementem tej pracy jest oddzielenie problemu reaktywności od selektywności. Ligand prowadzący do nieaktywnego katalizatora nie powinien być traktowany w tym samym modelu regresyjnym co ligandy dające produkt, dlatego najpierw wyznaczano obszar aktywnych

katalizatorów, a dopiero w jego obrębie poszukiwano zależności struktura–selektywność. Autorzy pokazali nawet, że dla reakcji Hayashiego–Hecka akceptowalny model regioselektywności można było uzyskać dopiero po wcześniejszym odfiltrowaniu ligandów na podstawie reaktywności. Praca ta pokazuje jak klasyfikację, regresję, analizę przestrzeni chemicznej i wirtualny screening można połączyć w praktyczne narzędzie wyboru liganda dla rzeczywistego problemu chemii procesowej, w którym poprawa jednej cechy katalizatora nie może odbywać się kosztem pozostałych.

4.5.2 Dobór katalizatora dla nowego substratu z wykorzystaniem danych literaturowych

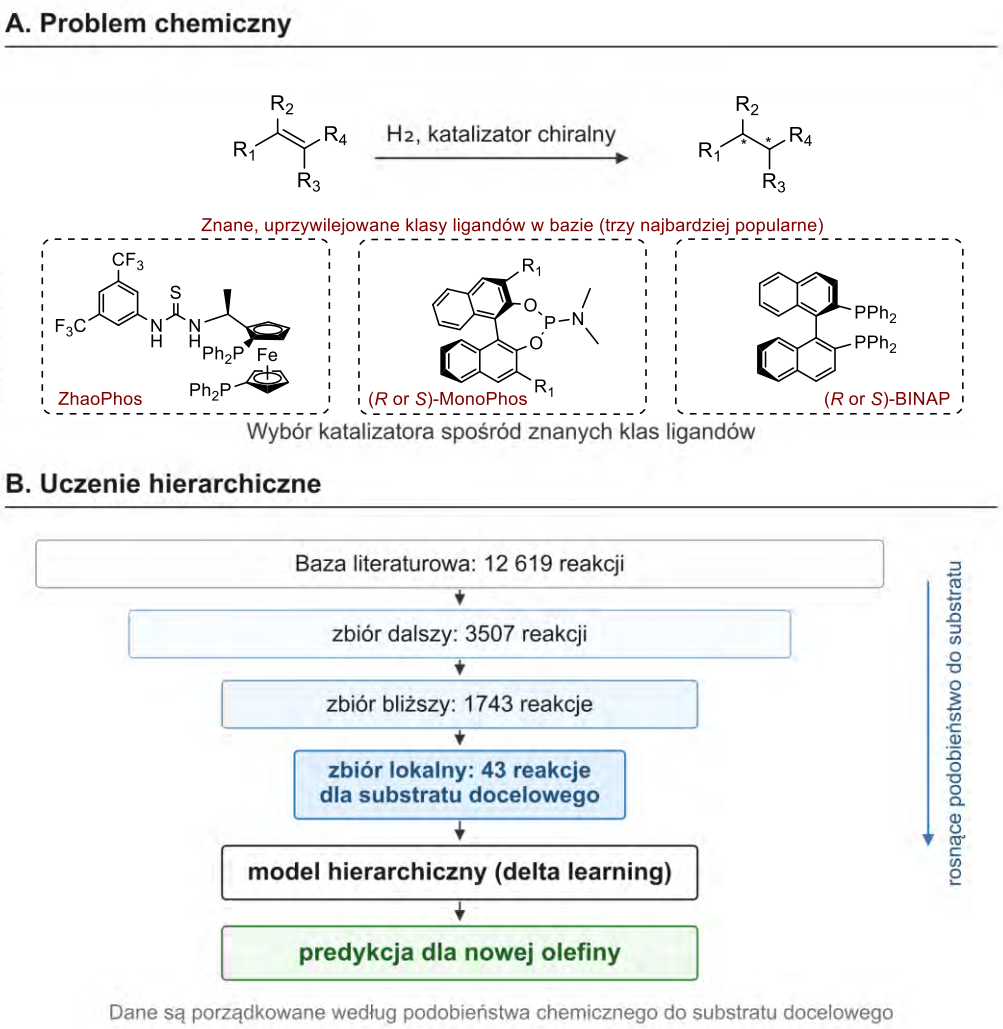
Szczególnie bliską praktyce laboratoryjnej odpowiedź na problem doboru katalizatora zaproponowali Xu i współpracownicy na przykładzie asymetrycznego uwodornienia olefin.[60] Jak schematycznie pokazano na Rysunku 15A, celem nie było projektowanie zupełnie nowego typu katalizatora, lecz wybór najlepszego spośród znanych klas ligandów dla konkretnej olefiny. Jest to sytuacja dobrze znana z praktyki: dla nowego substratu chemik zwykle dysponuje kilkoma wynikami wstępnego screeningu, natomiast dużo więcej danych istnieje już w literaturze dla substratów podobnych.

Punktem wyjścia była ręcznie komponowana baza 12 619 reakcji asymetrycznego uwodornienia olefin, zebranych z 355 publikacji. Obejmowała ona 2754 substraty olefinowe i 1686 chiralnych katalizatorów. Dla każdej reakcji zapisano strukturę substratu, katalizatora i produktu, warunki reakcji oraz wynik stereochemiczny. Taka baza nie została jednak potraktowana jako jeden jednorodny zbiór treningowy. Zamiast tego autorzy uporządkowali dane według bliskości chemicznej do substratu, co stanowi zasadę całego podejścia przedstawionego na Rysunku 15B.

Realizację zaproponowanej metody przeprowadzono dla wybranego substratu, dla którego dostępnych było 86 reakcji, z tej liczby 43 wykorzystano do treningu, a 43 do testowania, aby zasymulować realistyczną sytuację małej liczby danych. Oprócz tego wyodrębniono dwa większe zbiory reakcji dla substratów podobnych: 1743 reakcje dla olefin bardziej zbliżonych strukturalnie oraz 3507 reakcji dla olefin dalszych, ale nadal pokrewnych. Model uczono więc hierarchicznie: najpierw na danych bardziej ogólnych, a następnie jego przewidywania korygowano na poziomach coraz bliższych chemicznie substratowi docelowemu. W praktyce oznaczało to przejście od wiedzy ogólnej do wiedzy lokalnej.

Substraty i katalizatory opisano za pomocą deskryptorów 2D i 3D. Wśród testowanych reprezentacji najlepsze wyniki uzyskano dla połączenia MBTR (*many-body tensor representation*), opisującego strukturę trójwymiarową, z fingerprintami molekularnymi opisującymi budowę 2D. MBTR uwzględnia geometrię przestrzenną cząsteczki, w tym odległości i kąty między atomami, ale jest niezależny od tego, jak cała cząsteczka jest ustawiona w układzie współrzędnych. Model zbudowany wyłącznie na 43 lokalnych punktach był zbyt słaby, a proste dołączenie kilku tysięcy reakcji pokrewnych poprawiało wynik tylko częściowo. Dopiero ich hierarchiczna organizacja pozwoliła istotnie poprawić jakość predykcji.

Można stwierdzić, że taki model przypomina realny przebieg pracy eksperymentalnej. Gdy pojawia się nowa olefina, nie ma się do dyspozycji setek wyników dla tego konkretnego substratu. Najczęściej dostępnych jest jedynie kilkadziesiąt lub mniej obserwacji, a reszta wiedzy pochodzi z reakcji podobnych. Praca Xu pokazuje, że takie dane można wykorzystać skuteczniej, jeżeli nie traktuje się ich jako równoważnych, lecz nadaje im wagę wynikającą z podobieństwa chemicznego do aktualnego problemu.



Rysunek 15. *Schemat hierarchicznego wykorzystania danych literaturowych do doboru katalizatora dla nowej olefiny w asymetrycznym uwodornieniu.* ***(A)*** *Problem chemiczny: wybór katalizatora spośród znanych klas ligandów.* ***(B)*** *Organizacja danych według podobieństwa chemicznego do substratu docelowego i kolejne poziomy modelu hierarchicznego.*

4.5.3 Rekomendacja kompleksu metal–ligand dla nowej reakcji

Interesujące rozwinięcie problemu rekomendacji katalizatora przedstawili Tang, Liu, Feng, Hong i współpracownicy w pracy opublikowanej po ukazaniu się pierwszej publikacji wchodzącej w skład niniejszej rozprawy. Badania te dotyczyły asymetrycznych addycji Michaela katalizowanych kompleksami metali z chiralnymi ligandami *N,N*-dioksydowymi (Rysunek 16).[61] Jest to rodzina ligandów rozwijana przez zespół Xiaominga Fenga, tworząca układy katalityczne z różnorodnymi kwasami Lewisa (Rysunek 16A).

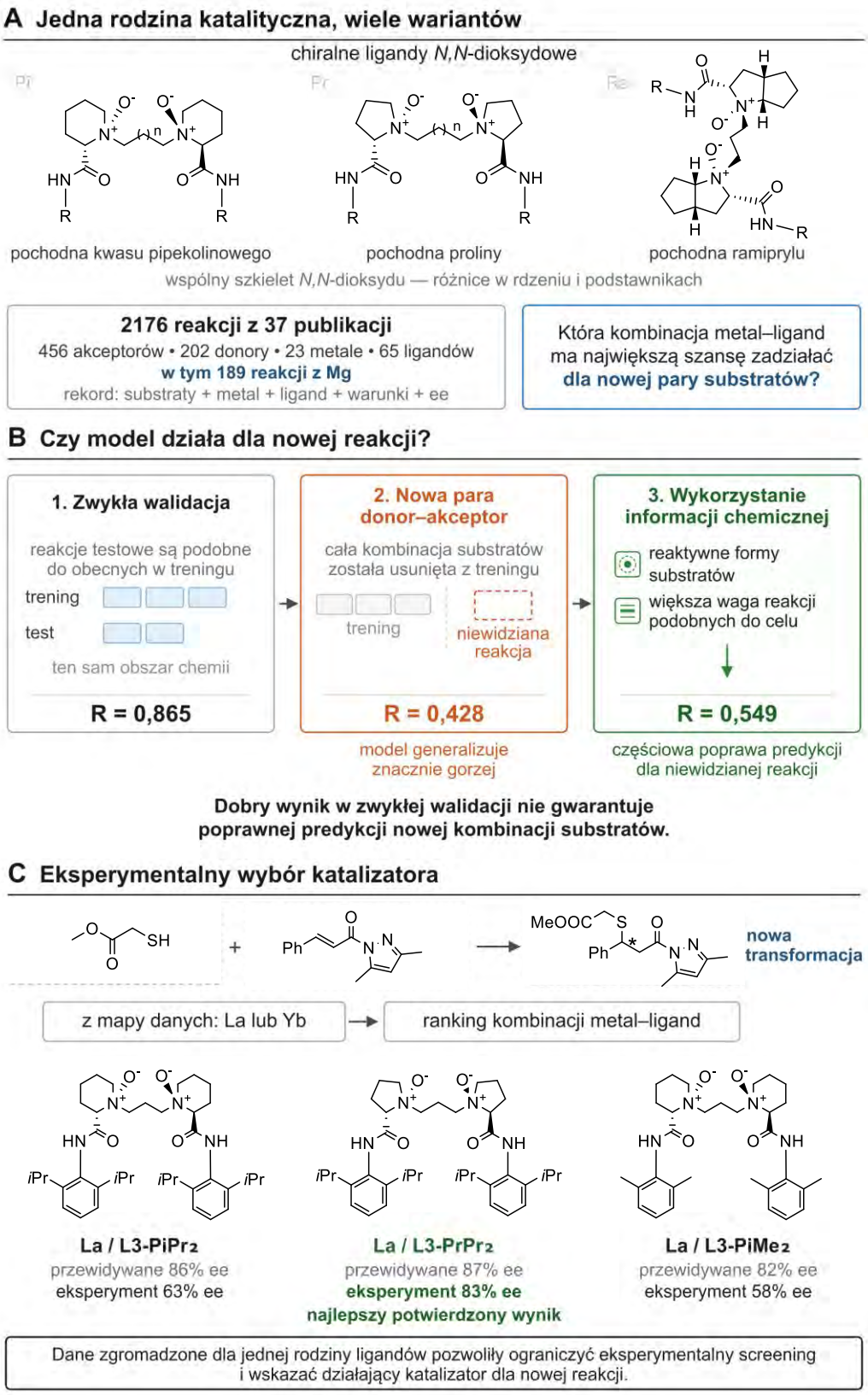
Autorzy wykorzystali rzeczywiste dane eksperymentalne zgromadzone podczas niemal dwóch dekad rozwoju tej rodziny katalizatorów. Z 37 publikacji i materiałów uzupełniających zebrano 2176 reakcji obejmujących 456 akceptorów, 202 donory, 23 metale i 65 ligandów. Dominowały metale ziem rzadkich, stanowiące 66,7% zbioru, przede wszystkim Sc, Yb, La i Y, ale licznie reprezentowane były również Mg, Co i Ni.

Najpierw autorzy wykorzystali dane do sporządzenia mapy stosowalności katalizatorów. Dla 1487 reakcji o *ee* ≥ 80% porównano, w jakich obszarach przestrzeni donorów i akceptorów poszczególne metale i rodziny ligandów dawały wysoką selektywność. Mapa pokazała, że skuteczność danego układu metal–ligand zależy od rodzaju substratów. Co więcej, nawet zmiana podstawników w obrębie tego samego szkieletu liganda może wyraźnie przesunąć obszar wysokiej enancjoselektywności. Taka mapa może więc pomóc zawęzić pierwszy screening do układów, które sprawdzały się wcześniej dla podobnych substratów.

Do przewidywania enancjoselektywności porównano kilka sposobów reprezentacji reakcji i modeli regresyjnych. Najlepsze wyniki uzyskała sieć CGRNN, która w standardowej pięciokrotnej walidacji osiągnęła Pearson R = 0,865 i MAE = 0,314 kcal mol^{-1}. Znacznie trudniejszy był jednak test, w którym z treningu całkowicie usuwano jedną parę donor–akceptor i model miał przewidywać katalizatory dla niewidzianej wcześniej transformacji (Rysunek 16B). Dla 17 takich przypadków, obejmujących 414 reakcji, korelacja spadła do R = 0,428. Pokazało to, że dobry wynik regresji przy losowym podziale danych nie oznacza jeszcze dobrej predykcji dla nowej reakcji. Diastereoselektywność potraktowano osobno: była dostępna aż dla 882 reakcji, czyli około 41% zbioru, i modelowano ją jako oddzielny cel. Dla modelu opartego na de uzyskano R^2 = 0,593. Nie była więc „mieszana" z głównym modelem enancjoselektywności, co stanowiłoby spore wyzwanie w sposobie zapisu danych i wprowadzenia ich do modelu.

Aby poprawić predykcję nowych transformacji, autorzy wprowadzili dwie zmiany. Po pierwsze, zamiast opisywać wyłącznie izolowane substraty, metal i ligand, uwzględnili reaktywną formę nukleofila oraz skoordynowany kompleks metal–ligand. Po drugie, podczas uczenia większą wagę nadawano reakcjom najbardziej podobnym do przewidywanego przypadku. Połączenie obu strategii zwiększyło średnią korelację w teście LORO z 0,428 do 0,549.

Praktyczną wartość podejścia sprawdzono dla nowej addycji tiolu do α,β-nienasyconego acylopirazolu (Rysunek 16C). Mapa stosowalności wskazała La i Yb jako najbardziej uzasadnione metale, a model uszeregował odpowiednie ligandy. Spośród trzech najwyżej ocenionych układów z lantanem najlepszy La/L3-$PrPr_2$ miał przewidywane 87% *ee*, a eksperymentalnie dał 83% *ee*. Autorzy pokazali więc, że dane historyczne mogą służyć nie tylko do przewidywania wartości *ee*, ale do zawężenia przestrzeni eksperymentalnego screeningu dla nowej reakcji.



Rysunek 16. *Wykorzystanie danych literaturowych do rekomendacji katalizatora w asymetrycznych addycjach Michaela katalizowanych kompleksami metali z ligandami N,N-dioksydowymi:* ***(A)*** *trzy główne rodziny ligandów i zakres zgromadzonej bazy danych,* ***(B)*** *ocena modelu dla niewidzianych par donor–akceptor oraz strategie poprawiające predykcję,* ***(C)*** *wybór i eksperymentalna walidacja kompleksów La/N,N-dioksyd dla nowej addycji Michaela.*

4.6 Generalizacja — czy model uczy się zależności chemicznych, czy tylko odtwarza znane dane?

4.6.1 Meta-uczenie — wykorzystanie podobnych reakcji przy małej liczbie danych

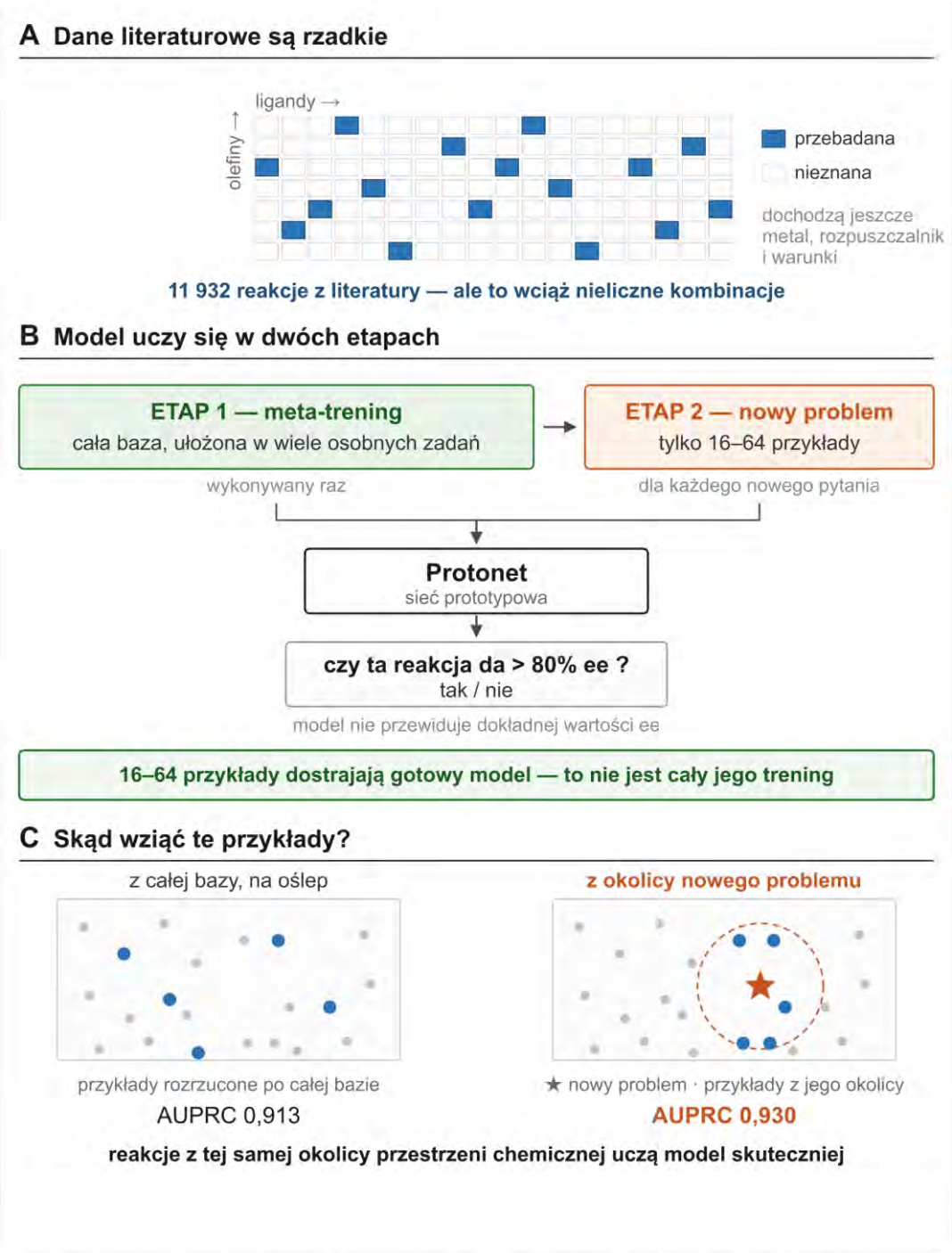
Singh i Hernández-Lobato powrócili do tej samej bazy literaturowych asymetrycznych uwodornień olefin, która stanowiła podstawę modelu Xu i współpracowników opisanego w podrozdziale 4.5.2, stawiając jednak inne pytanie: nie jak zbudować lokalny model dla jednego substratu, lecz czy można nauczyć model wykorzystywać wiedzę z wielu pokrewnych reakcji wtedy, gdy dla nowego problemu dostępnych jest tylko kilka przykładów.[62] Z pierwotnej bazy wybrano reakcje katalizowane irydem, rodem i kobaltem, a słabo reprezentowane dane kobaltowe uzupełniono o nowsze przykłady literaturowe; ostateczny zbiór obejmował 11 932 reakcje, w tym 5009 irydowych, 6391 rodowych i 532 kobaltowe. W bazie wyjściowej około 90% reakcji dotyczyło rodu i irydu, a kobalt stanowił około 1%, co dobrze pokazuje nierównomierną reprezentację układów katalitycznych w danych literaturowych.

Zadanie sformułowano jako pytanie, które mogłoby paść w praktyce: czy dana kombinacja substratu, katalizatora i warunków prowadzi do enancjoselektywności większej niż 80% *ee* (Rysunek 17B). Model nie przewidywał zatem dokładnej wartości *ee*, lecz rozpoznawał kombinacje rokujące, a jako główną miarę jakości przyjęto AUPRC (*Area Under the Precision–Recall Curve* czyli: pole pod krzywą precyzja–czułość.), odpowiednią dla niezrównoważonych liczebności klas (jedna klasa w stopniu znacznym przeważa). Jest to rodzaj klasyfikacji, wykorzystany do odnajdywania reakcji rokujących (powyżej 80% *ee*) w przeważającej liczbie reakcji nierokujących (poniżej 80% *ee*).

Następnie zastosowano sieci prototypowe (Protonet) należące do metod meta-uczenia: model uczył się najpierw wspólnych prawidłowości na wielu zadaniach utworzonych z całej bazy, a dopiero potem dostrajano go do nowego zadania przy użyciu 16, 32 lub 64 przykładów pomocniczych. Meta-uczenie było tam potrzebne dlatego, że dla nowego problemu chemicznego autorzy zakładali dostęp tylko do bardzo małej liczby nowych eksperymentów (16, 32 albo 64), a chcieli mimo to rozpoznać obiecujące kombinacje.

Szczególnie istotny okazał się dobór danych pomocniczych. Gdy przestrzeń reakcji pogrupowano według podobieństwa substratów, ligandów i warunków, a do przewidywania wykorzystywano reakcje z tego samego obszaru przestrzeni chemicznej, AUPRC wzrosło do 0,930. W procedurze *leave-one-cluster-out*, w której jeden z dziesięciu obszarów całkowicie wyłączano z treningu, Protonet przewyższał modele jednozadaniowe we wszystkich klastrach, choć dla dwóch z nich AUPRC pozostawało poniżej 0,80. Na niezależnym zbiorze 245 nowszych reakcji irydowych i rodowych, nieobecnych w bazie pierwotnej, Protonet osiągnął przy 64 przykładach pomocniczych AUPRC 0,934, wobec 0,863 dla lasu losowego i 0,876 dla sieci grafowej.

Z punktu widzenia chemika najważniejszy jest wariant niewymagający wykonywania dodatkowych eksperymentów: przykłady pomocnicze wybierano spośród już istniejących danych treningowych na podstawie ich podobieństwa do przewidywanego przypadku, uzyskując dla zbioru zewnętrznego AUPRC 0,915. Ta sama duża i niepełna baza literaturowa, którą wcześniej wykorzystano do zbudowania lokalnego modelu dla konkretnego substratu, może więc posłużyć również do nauczenia modelu ogólnej strategii korzystania z wcześniejszych obserwacji.



Rysunek 17. *Meta-uczenie na rzadkich danych literaturowych. (A) Tylko niewielka część możliwych kombinacji składników została zbadana. (B) Model przechodzi meta-trening na całej bazie, a dopiero potem dostraja się do nowego problemu przy użyciu 16–64 przykładów. (C) Przykłady pomocnicze dobrane spośród reakcji podobnych do nowego przypadku działają lepiej niż wybrane z całej bazy.*

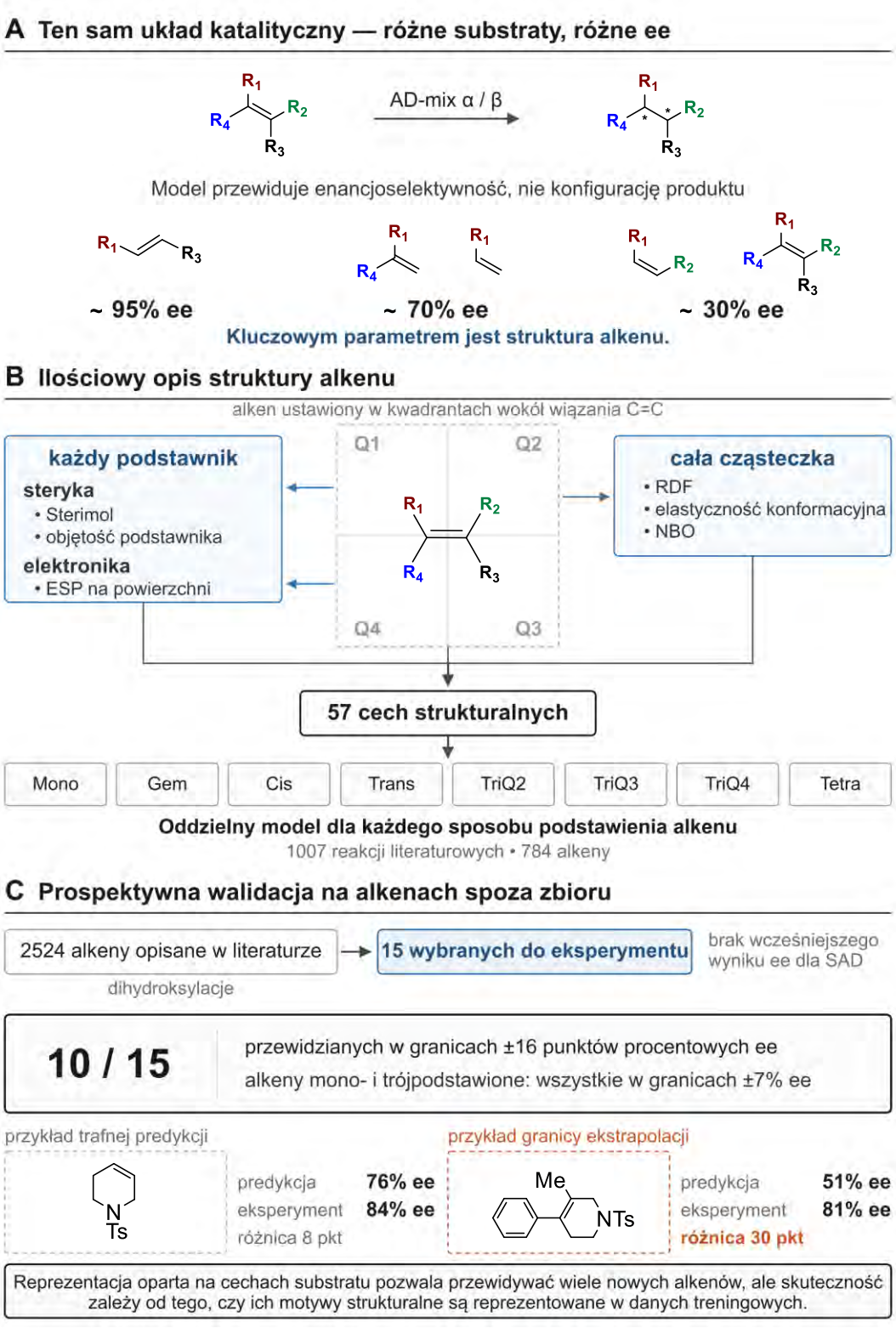
4.6.2 Decydujący wpływ struktury substratu na enancjoselektywność

Ocampo, Altundas, Denmark i współpracownicy zbadali generalizację modelu w asymetrycznej dihydroksylacji Sharplessa (SAD), w której dla różnych alkenów stosuje się standardowe układy katalityczne zawierające chiralne ligandy pochodne alkaloidów chinowca (AD-mix α lub β).[63] W tym przypadku głównym źródłem zmienności enancjoselektywności jest więc struktura substratu. Autorzy postawili pytanie, czy na podstawie budowy nowego alkenu można przewidzieć, jak wysokiej enancjoselektywności należy oczekiwać w standardowych warunkach SAD.

Z literatury zebrano 1007 reakcji obejmujących 784 alkeny. Do modelowania wykorzystano prochiralne substraty, a przewidywanym wynikiem była wielkość enancjoselektywności, niezależnie od tego, który enancjomer powstawał. Alkeny podzielono na osiem klas według sposobu podstawienia wiązania C=C. Najbardziej chemicznym elementem pracy było przekształcenie klasycznego modelu kwadrantowego Sharplessa–Norrby'ego w opis liczbowy: podstawniki umieszczano w czterech kwadrantach wokół wiązania podwójnego i opisywano ich właściwości steryczne i elektronowe oraz geometrię całej cząsteczki (Rysunek 18A–B).

Najważniejszym testem była walidacja na nowych substratach. Spośród ponad 2500 alkenów opisanych wcześniej w innych reakcjach dihydroksylacji z użyciem osmu autorzy wybrali 15 struktur, dla których nie raportowano wcześniej enancjoselektywności w warunkach SAD (Rysunek 18C). Dla 10 z 15 przypadków błąd predykcji nie przekraczał 16 punktów procentowych *ee*, a dla wszystkich badanych alkenów monosubstytuowanych i trójpodstawionych wynosił najwyżej około 7% *ee*. Jednocześnie model wyraźnie zawiódł dla jednego alkenu czteropodstawionego, dla którego przewidywano 51% *ee*, a eksperymentalnie otrzymano 81% *ee*.

Praca pokazuje więc generalizację do nowych substratów przy zachowaniu tego samego układu katalitycznego. Jej znaczenie polega nie tylko na przewidywaniu *ee*, lecz także na wskazaniu, dla których klas alkenów model potrafi przenieść wiedzę poza dane treningowe, a gdzie taka ekstrapolacja staje się zawodna. Analiza cech modelu dodatkowo sugerowała, że względne znaczenie efektów sterycznych, elektronowych i konformacyjnych zmienia się pomiędzy klasami alkenów, ale autorzy traktowali te zależności jako źródło hipotez, a nie dowód mechanizmu.



Rysunek 18. *Przewidywanie enancjoselektywności dihydroksylacji Sharplessa na podstawie struktury substratu: (A) zależność wielkości ee od budowy alkenu przy zastosowaniu standardowych układów AD-mix α/β, (B) ilościowa reprezentacja alkenu oparta na modelu kwadrantowym Sharplessa–Norrby'ego oraz (C) eksperymentalna walidacja modeli na 15 alkenach bez wcześniejszych danych o enancjoselektywności w warunkach SAD.*

4.6.3 Model uczy się reaktywności czy literatury?

Jeszcze inne ograniczenie modeli uczonych na danych literaturowych pokazali Beker, Grzybowski i współpracownicy na przykładzie heterocyklicznych sprzęgań Suzuki–Miyaury.[64] Na podstawie starannie opracowanego zbioru ponad 10 000 reakcji literaturowych próbowano przewidywać, jakie warunki należy zastosować dla nowej pary reagentów, nawet po ograniczeniu zadania jedynie do wyboru zasady i rozpuszczalnika. Niezależnie od zastosowanego algorytmu i sposobu reprezentacji struktur modele nie przewidywały warunków istotnie lepiej niż prosta strategia wybierająca te, które najczęściej występowały w literaturze. Autorzy wskazali, że model może w takiej sytuacji uczyć się przede wszystkim utrwalonych praktyk eksperymentalnych, czyli preferencji dotyczących popularnych protokołów, dostępności odczynników oraz efektu braku nieudanych reakcji, zamiast rzeczywistej zależności między strukturą reagentów a optymalnymi warunkami. Warto odnotować, że praca ta bywa później przywoływana jako przykład skutecznego modelowania złożonych danych reakcyjnych,[65] mimo że jednym z jej głównych wniosków jest właśnie ostrzeżenie przed taką interpretacją wysokiej jakości predykcji.

Wniosek uzupełnia problem omówiony wcześniej w podrozdziale 3.4.3 przy komentarzu Chuanga i Keisera.[40] Tam podobny wynik uzyskiwany z cech chemicznych i nieinformatywnych wskazywał, że sposób konstrukcji zbioru i walidacji może zawyżać ocenę modelu. W pracy Bekera problem leżał natomiast w samych danych literaturowych: duża liczba rekordów nie gwarantowała, że zawierają one informację potrzebną do odtworzenia reaktywności. Oba przykłady prowadzą więc do tego samego zastrzeżenia istotnego dla generalizacji: dobra predykcja wzorców obecnych w bazie nie jest jeszcze dowodem, że model nauczył się zależności chemicznej.

4.6.4 Co ujawnia zaostrzenie kryteriów oceny modeli

Znaczenie sposobu podziału danych pokazali szczególnie wyraźnie Lee, Moldagulov i Grzybowski w badaniu przewidywania bioaktywności.[66] Autorzy zadali pytanie istotne również dla modeli reakcji chemicznych: czy dobra jakość predykcji przy podziale losowym oznacza rzeczywistą zdolność przewidywania nowej chemii, czy jedynie dobrą interpolację między podobnymi strukturami.

Porównano sześć architektur modeli na dziesięciu zbiorach ChEMBL, stosując podziały na zestaw treningowy i testowy wedle rosnącego stopnia rozróżnienia strukturalnego: losowy, według rdzeni Bemisa–Murcko, według klastrów Butiny, według zbioru zróżnicowanego oraz wedle UMAP i odcięcia chronologicznego (prawdziwy test na prospektywność). Wraz ze wzrostem trudności podziału skuteczność wszystkich modeli malała: w klasyfikacji z około 80% zbalansowanej dokładności przy podziale losowym do około 58% przy testach najbardziej wymagających. Dodanie deskryptorów trójwymiarowych nie zmieniało tego obrazu.

Kluczowe było jednak bezpośrednie porównanie podobieństwa struktur po obu stronach podziału. Przy podziale losowym, a także przy podziale według rdzenia Bemisa–Murcko, cząsteczki testowe pozostawały często bardzo podobne do treningowych, natomiast przy podziałach UMAP i chronologicznym podobieństwo to było znacznie mniejsze. Jakość predykcji okazała się przy tym niemal liniową funkcją owego podobieństwa ($R^2 = 0{,}98$): im bardziej cząsteczka testowa przypominała dane treningowe, tym lepiej działał model. Prosty model najbliższych sąsiadów dorównywał przy tym architekturom znacznie bardziej złożonym.

Wniosek jest bezpośrednio istotny dla chemii: wysoka dokładność przy podziale losowym może ukrywać fakt, że model korzysta głównie z podobieństwa do przykładów treningowych. Do

oceny generalizacji potrzebne są zatem podziały bardziej wymagające oraz proste punkty odniesienia, takie jak model najbliższych sąsiadów, pokazujące, ile da się osiągnąć wyłącznie na podstawie podobieństwa strukturalnego.

Praca ta dotyczy przewidywania bioaktywności, a nie reaktywności ani selektywności, a odległość od zbioru treningowego zdefiniowano w niej wyłącznie strukturalnie, w przestrzeni odcisków molekularnych. Ukazała się ponadto w 2026 roku, a więc po opublikowaniu pracy stanowiącej podstawę rozdziału 6.1.

5. Podsumowanie stanu wiedzy

Przegląd przedstawiony w rozdziałach 3 i 4 pokazał, że uczenie maszynowe może pełnić w katalizie asymetrycznej różne funkcje, od przewidywania enancjoselektywności po wybór katalizatora i analizę jego zakresu stosowalności. Celem tego rozdziału jest uporządkowanie stanu wiedzy wokół dwóch pytań, na które odpowiadają publikacje wchodzące w skład rozprawy. Pierwsze dotyczy wykorzystania rozproszonych danych literaturowych do rekomendowania katalizatora. Drugie dotyczy tego, jakie pytania predykcyjne można wiarygodnie stawiać małym zbiorom powstającym podczas rozwoju pojedynczej reakcji. Oba prowadzą również do wspólnego problemu generalizacji i eksperymentalnej weryfikacji predykcji.

5.1 Stan wiedzy w momencie rozpoczęcia badań

Do końca 2023 roku uczenie maszynowe było już stosowane w katalizie asymetrycznej do przewidywania enancjoselektywności, wyboru katalizatora z wcześniej określonej przestrzeni struktur oraz analizy zależności pomiędzy jego budową a wynikiem reakcji. Podejścia te działały jednak najczęściej w obrębie jednej dobrze zdefiniowanej transformacji lub jednej rodziny katalizatorów.

Dostępne dane koncentrowały się przy tym wokół najlepiej rozwiniętych układów, przede wszystkim kompleksów metali przejściowych oraz chiralnych kwasów fosforowych. Skalę tej nierównowagi dobrze ilustruje późniejsze zestawienie asymetrycznych uwodornień olefin, obejmujące 6391 reakcji katalizowanych rodem, 5009 irydem i jedynie 532 kobaltem.[62] Pokazuje to, że nawet w obrębie jednej dobrze poznanej transformacji dostępność danych może być silnie zależna od popularności danego metalu.

5.2 Problem badawczy I. Czy rozproszona literatura może służyć do rekomendowania katalizatora?

Problemem jest nie tylko przewaga wyników pozytywnych nad negatywnymi, lecz także nadreprezentacja reakcji cieszących się większym zainteresowaniem badaczy. Beker i współpracownicy pokazali, że model uczony na danych literaturowych może odzwierciedlać popularność określonych reakcji w publikacjach zamiast wskazywać realne trendy w danych. [64] Wybór zbioru danych wpływa więc nie tylko na jakość predykcji, lecz także na to, które obszary chemii w ogóle stają się dostępne dla modelowania.

Magnez był w tym obrazie niemal nieobecny. Nie wynika to z braku rozwiniętej chemii asymetrycznej tego metalu, której zakres omówiłam w rozdziale 2, lecz z rozproszenia dostępnych wyników pomiędzy wieloma typami reakcji i układami katalitycznymi. Wśród prac dostępnych do końca 2023 roku nie znalazłam modelu przewidującego enancjoselektywność dla układów magnezowych. Podobnie niewiele uwagi poświęcano innym metalom rzadziej wykorzystywanym w katalizie asymetrycznej.

Dotychczasowe narzędzia służyły głównie do optymalizacji układu w obrębie jednej znanej reakcji. Otwarte pozostawało pytanie, czy z niejednorodnej bazy obejmującej wiele typów reakcji jednego metalu można zbudować użyteczną rekomendację katalizatora dla nowego problemu syntetycznego. Taki zbiór nie ma wspólnego rdzenia strukturalnego ani jednolitego protokołu eksperymentalnego, co istotnie odróżnia go od większości wcześniej modelowanych układów. Odpowiedzi na to pytanie udziela rozdział 6.1.

5.3 Problem badawczy II. Jakie pytania można stawiać małym zbiorom laboratoryjnym?

Sama niewielka liczebność zbioru nie wyklucza modelowania. W literaturze istniały już przykłady modeli budowanych na kilkudziesięciu punktach, jeżeli badany problem był dobrze zdefiniowany, a dane wewnętrznie spójne. Sama liczba obserwacji nie rozstrzyga jednak, jakie pytanie predykcyjne można takiemu zbiorowi wiarygodnie postawić.

Szczególnym przypadkiem są dane powstające w etapie następującym po optymalizacji metody katalitycznej, czyli podczas badania zakresu stosowalności metody na wybranych substratach. Nie są one projektowane z myślą o modelowaniu, lecz mają wykazać stopień uniwersalności i ograniczenia opracowanej metody. Nie było więc jasne, kiedy taki mały, eksperymentalnie jednorodny zbiór pozwala przewidywać wartość *ee*, a kiedy dostępna informacja wystarcza jedynie do rozstrzygania reaktywności lub porządkowania substratów według oczekiwanego wyniku.

Problem ten jest bezpośrednio związany z niedoborem danych negatywnych. Zakresy substratowe są z natury przesunięte w stronę wyników udanych, a przydałoby się również dogłębniej poznać ograniczenia metody. Jednocześnie reakcja o niskiej enancjoselektywności, reakcja o niskiej wydajności i reakcja, w której produkt nie powstaje, niosą różne informacje chemiczne i nie powinny być traktowane jako jeden typ wyniku. Model uczony wyłącznie na opublikowanych sukcesach nie otrzymuje pełnej informacji o granicach reaktywności badanego układu. Odpowiedzi na pytanie, jakie predykcje można wiarygodnie uzyskać z takich danych, udziela rozdział 6.2.

5.4 Wspólne ograniczenie obu podejść. Generalizacja i walidacja eksperymentalna

Oba ograniczenia prowadzą do tego samego problemu. Model może zwrócić predykcję także dla struktury słabo reprezentowanej lub całkowicie nieobecnej w danych treningowych, a sama wartość liczbowa nie informuje, czy jest to interpolacja w obrębie znanej przestrzeni chemicznej, czy ekstrapolacja poza nią. Z tego względu sposób podziału danych i analiza zakresu stosowalności są niezbędne do oceny wiarygodności predykcji.

Walidacja na wydzielonym zbiorze testowym nadal opiera się jednak na danych już istniejących. Prospektywna walidacja eksperymentalna pozwala sprawdzić model na przypadkach, których wynik nie był znany w momencie formułowania predykcji, dlatego stanowi najbardziej bezpośredni test jego praktycznej użyteczności. Wśród prac omawianych w rozdziale 4 takie eksperymentalne sprawdzenie predykcji nadal należało do wyjątków, podczas gdy w większości przypadków wyniki modelu porównywano z wcześniej opublikowanymi danymi.

W obu pracach wchodzących w skład niniejszej rozprawy zdolność predykcyjną modeli sprawdzałam eksperymentalnie na nowych przypadkach nieobecnych w danych treningowych. Szczegółowy opis tych walidacji przedstawiono w rozdziałach 6.1 i 6.2.

5.5 Późniejsza literatura jako potwierdzenie kierunku

Niektóre prace przywołane w tym podrozdziale ukazały się już po publikacji stanowiącej podstawę rozdziału 6.1 i nie mogły stanowić motywacji dla rozpoczęcia tych badań. Są jednak istotne, ponieważ rozwijają te same problemy, które wskazano powyżej jako otwarte.

Gallarati, Bucci, Doyle i Sigman pokazali możliwość przenoszenia informacji pomiędzy pokrewnymi reakcjami przy ograniczonej liczbie danych.[55] Singh i Hernández Lobato zastosowali meta uczenie do adaptacji modelu do nowych zadań na podstawie niewielkiej liczby przykładów.[62] Ocampo i Denmark przeprowadzili prospektywną walidację modelu dla asymetrycznej dihydroksylacji Sharplessa,[63] natomiast Gao i Coley analizowali wpływ skrzywienia publikacyjnego na wnioski dotyczące reaktywności.[67] Pod koniec 2025 roku opublikowano również wytyczne dobrych praktyk stosowania uczenia maszynowego w katalizie homogenicznej.[26] Prace te nie wyznaczyły kierunku niniejszej rozprawy, ale pokazują, że problemy generalizacji, jakości danych i wiarygodnej walidacji zostały następnie szerzej rozpoznane w tej dziedzinie.

Szczegółowy kontekst projektów i badania własne

6.1 Rekomendacja układu katalitycznego na podstawie danych literaturowych

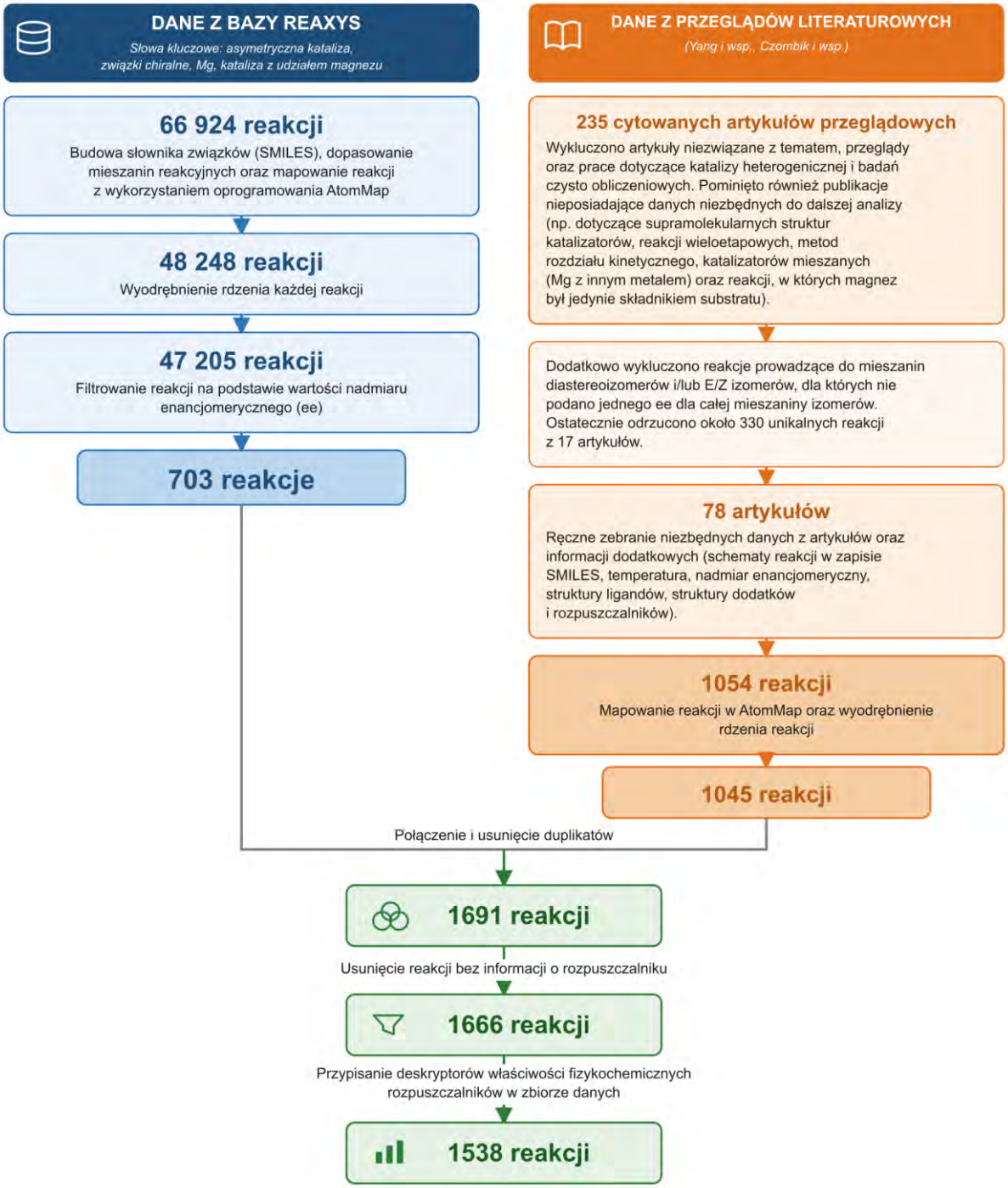
Celem tej części pracy było zbudowanie modelu opartego na danych literaturowych, który zautomatyzuje pierwszy etap optymalizacji reakcji katalizowanych magnezem i wskaże układ katalityczny możliwy do sprawdzenia w obrębie pierwszych pięciu propozycji. Dotyczy to dwóch sytuacji: poszukiwania alternatywnych warunków dla reakcji już opisanej oraz doboru chiralnego liganda i soli magnezu dla reakcji nowej. Oprócz tego jak często model trafia w odpowiedni ligand na pierwszej pozycji rankingu, interesowało mnie czy kolejne pozycje stanowią realną alternatywę, bo dopiero to otwiera drogę do zamiany liganda na tańszy, trwalszy albo po prostu dostępny w szafie laboratoryjnej.

Kataliza magnezowa jest dla takiego zadania przypadkiem reprezentatywnym i zarazem trudnym. Reprezentatywnym, ponieważ magnez należy do metali pospolitych i stanowi realną alternatywę dla metali o ograniczonych zasobach, natomiast trudnym, ponieważ układy magnezowe są w zbiorach danych wykorzystywanych w uczeniu maszynowym niedostatecznie reprezentowane. Skalę tego pominięcia dobrze oddaje obserwacja z przeglądu w rozdziale 4: ten sam zbiór 1075 reakcji addycji tioli do imin katalizowanej chiralnymi kwasami fosforowymi, opracowany w zespole Denmarka, powraca jako punkt odniesienia w co najmniej czterech niezależnych pracach, podczas gdy dla magnezu nie powstał ani jeden model przewidujący enancjoselektywność. Warto też wspomnieć, że te 1075 reakcji zawiera jeden typ reakcji.

Implementację sieci neuronowej, operacje na bazie danych i aplikację sieciową wykonywał dr M. Kulczykowski, a we wcześniejszych etapach prac nad modelem uczestniczył dr R. Roszak. Nad uzupełnianiem danych pracowali ze mną dr hab. B. Zambroń i dr J. Jaszczewska-Adamczak. Poniżej opisuję pracę w kolejności, w jakiej przebiegała, wskazując przy każdym etapie, co wykonałam samodzielnie, a co powstało we współpracy. Szczegółowy podział wkładu przedstawiają oświadczenia współautorów zamieszczone na końcu rozprawy.

6.1.1 Budowa zbioru danych

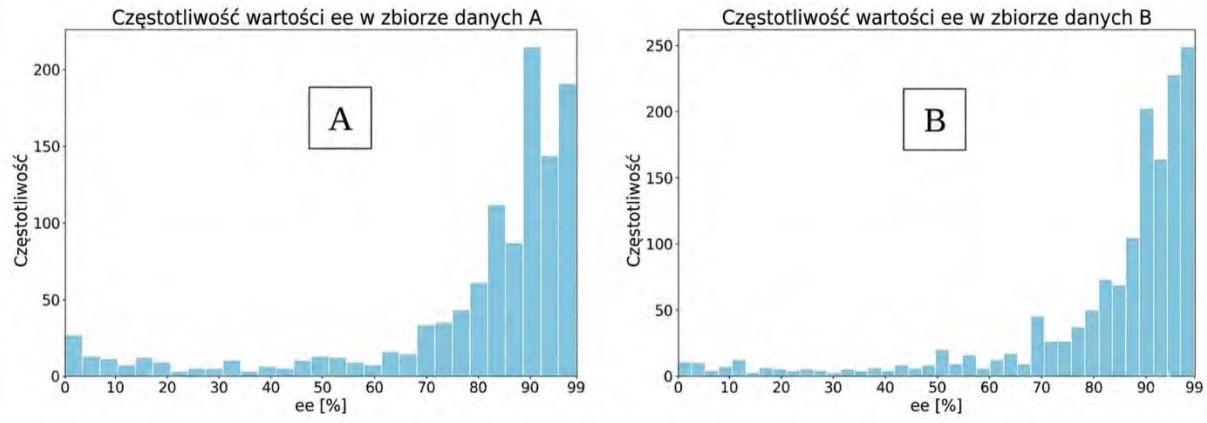
Zbudowanie wiarygodnego zbioru danych okazało się najtrudniejszym i najbardziej czasochłonnym elementem całego projektu. Uzyskany zbiór bezpośrednio opowiada otrzymanej jakości predykcji. Pracę rozpoczęłam od systematycznego przeszukania repozytorium Reaxys hasłami „asymmetric catalysis”, „chiral”, „Mg” oraz „magnesium-catalyzed”, uzyskując 66 924 rekordy. Znaczna ich część nie dotyczyła jednak reakcji stereoselektywnych albo nie zawierała informacji niezbędnych do modelowania. Po odfiltrowaniu wpisów bez przypisanej wartości nadmiaru enancjomerycznego oraz bez kompletnych warunków katalitycznych pozostały 703 reakcje. Spadek o dwa rzędy wielkości jest tu wynikiem samym w sobie: pokazuje, ile z zapisanej wiedzy chemicznej pozostaje niedostępne dla automatycznej ekstrakcji, mimo że w publikacjach źródłowych jest obecne. Schemat procesu zbierania danych przedstawia Rysunek 19.



***Rysunek 19**. Budowa zbioru danych dla katalizy magnezowej: ekstrakcja automatyczna wobec opracowania ręcznego, dwie decyzje kuratorskie i zbiór końcowy.*

Zbiór tej wielkości nie wystarczał, dlatego rozszerzyłam go opracowaniem ręcznym. Zaprojektowałam procedurę zbierania danych — co i skąd jest pobierane, w jakiej postaci zapisywane i według jakich kryteriów odrzucane — a następnie nadzorowałam jej przebieg, wielokrotnie zmieniając przyjęte normy w miarę napotykania problemów, których na początku nie dało się przewidzieć. Wraz z dr. hab. B. Zambroniem i dr J. Jaszczewską-Adamczak przejrzeliśmy 235 artykułów, wybraliśmy 78 artykułów źródłowych wraz z materiałami uzupełniającymi, odzyskując 1054 reakcje.

Największą wartością tego przeglądu był dostęp do badań optymalizacyjnych opisanych w materiałach uzupełniających. Dostarczyły one nie tylko reakcji o wysokiej selektywności, lecz również wyników niskich i negatywnych, które w ustrukturyzowanych bazach danych praktycznie nie występują — odzyskaliśmy w ten sposób 128 przykładów o nadmiarze enancjomerycznym poniżej 50%. Na tle całej bazy to nadal nie jest satysfakcjonująca ilość, ale trzeba przyjąć, że jest to cecha zbioru, a nie metody komponowania zestawu danych (Rysunek 20).



Rysunek 20. *Histogramy wartości ee w zbiorze opracowanym ręcznie (A) i w zbiorze pozyskanym automatycznie (B).*

Dwie decyzje podjęte na tym etapie warto opisać osobno, ponieważ mają bezpośredni wpływ na to, czego model może się nauczyć oraz pokazują wyzwania obecne podczas budowy zbioru danych.

Pierwsza dotyczyła mieszanin diastereoizomerów i okazała się problemem nieoczekiwanym. Sposób raportowania takich reakcji jest w literaturze niejednorodny: raz podaje się jeden wynik dla całej mieszaniny, innym razem rozdziela się stereoizomery i przypisuje im osobne wartości *ee*, ale wydajność pozostaje podana dla mieszaniny. Dla chemika czytającego pracę jest to zapis akceptowalny i poprawny, ale dla zbioru uczącego jest nie do przyjęcia, ponieważ modelowanie wymaga jednorodności rekordu. Jeżeli każda reakcja ma przypisane *ee*, wydajność i typ, to nie może istnieć podzbiór reakcji z wartością *ee*, lecz bez wydajności, bo wydajność dotyczy mieszaniny, a poszczególnych produktów nie izolowano. Zdecydowałam o usunięciu takich wpisów. Jest to problem istotny dla całej dziedziny: przyjęta i całkowicie poprawna praktyka raportowania uniemożliwia wykorzystanie sporego zakresu opublikowanych danych. Podobny problem występował w pracy opublikowanej dwa lata później przez Xiominga Fenga opisanej w podrozdziale 4.5.3. Tam decyzja o włączeniu reakcji diastereoselektywnych polegała na utworzeniu osobnego modelu dla tych reakcji. Gdybyśmy mieli rozwiązać ten problem, to postąpilibyśmy podobnie, jednak ilość samych reakcji produkujących diastereoizomery była niewystarczająca do wytrenowania analogicznego modelu – sieci neuronowej, które wymagają większej ilości danych ze względu na tendencje do przeuczenia. Musielibyśmy skomponować odrębny nowy model – bez analogii architektury, dlatego uznaliśmy, że skupimy się na reakcjach enancjoselektywnych.

Druga dotyczyła nazewnictwa reagentów. Nazwy katalizatorów, ligandów i dodatków zapisywano w bazie według niejednolitych konwencji, np. ta sama struktura występuje raz jako (1R,1′R,2S,2′S)-1,1′-(propano-1,3-diylo)bis(2-((2,6-diizopropylofenylo)karbamoilo)pirolidyno-1-tlenek), a raz jako wzór sumaryczny $C_{37}H_{56}N_4O_4$. Poprosiłam dr. Kulczykowskiego o wyekstrahowanie wszystkich nazw reagentów występujących w zbiorze, a następnie pozycja po pozycji, ręcznie, przypisałam każdej z nich zapis SMILES. Dopiero wtedy było wiadomo, ile

różnych nazw może odpowiadać jednemu reagentowi. Powstały słownik posłużył do automatycznego uzupełnienia całej bazy.

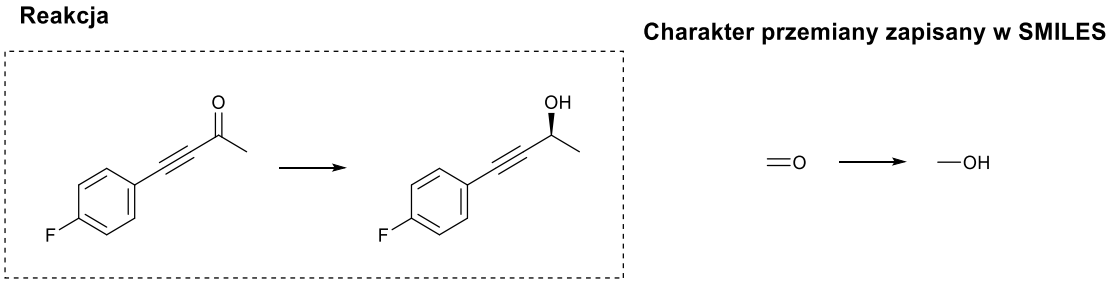
Na tej podstawie ustaliłam normę wyodrębniania z pola reagentów trzech osobnych składników: chiralnego liganda, źródła magnezu oraz dodatku wpływającego na indukcję asymetryczną. Wyodrębnienie dodatków wymagało sięgnięcia do publikacji źródłowej wszędzie tam, gdzie obok soli magnezu i liganda występował jeszcze jakiś reagent. Pełną listę dodatków opracowaliśmy wspólnie z dr. hab. B. Zambroniem i dr J. Jaszczewską-Adamczak. Dodatki mogące wpływać na strukturę katalizatora, równowagę reakcji lub mechanizm[68, 69], np. kwas 4-nitrobenzoesowy albo alkohol benzylowy - występują w około 18% reakcji zbioru.

Scalenie obu zbiorów, usunięcie duplikatów i zapis w bazie SQLite wykonał dr Kulczykowski. Duplikaty identyfikowano przez kanonizację zapisów SMILES z użyciem biblioteki MolVS[70]: ciąg dzielono znakiem „>>” na stronę substratową i produktową, standaryzowano obie strony osobno i łączono ponownie. Niepowodzenie tej procedury służyło jako pierwszy filtr wychwytujący reakcje zapisane niepoprawnie. Odczyt wartości *ee* i temperatury z niejednorodnych zapisów - „98 *ee*”, „98 percent *ee*”, wartości umieszczone w kolumnie komentarza zamiast w polu wyniku - wymagał osobnych rozwiązań, które opracowywałam na bieżąco, a wykonywał je na bazie danych dr Kulczykowski. Wszystkie potencjalne błędy i reakcje, które nie przechodziły przez procedury oraz decyzja co należy z tym zrobić leżały po mojej stronie.

Po usunięciu duplikatów i weryfikacji dostępności deskryptorów zbiór końcowy objął 1538 reakcji, 139 chiralnych ligandów i 104 typy reakcji. Mimo tej skali ma on istotne ograniczenie: rozkład wartości *ee* pozostaje wyraźnie przesunięty ku wysokim selektywnościom, co odzwierciedla uprzywilejowanie wyników udanych w publikacjach. Udział 128 odzyskanych reakcji o niskim *ee* tego nie równoważy (Rysunek 20). Pełny zbiór udostępniono publicznie pod adresem http://asymmetric.icho.edu.pl/.

6.1.2 Reprezentacja numeryczna reakcji i warunków

Reakcję i warunki jej prowadzenia opisano wektorem liczbowym złożonym z kilku niezależnych bloków. Substraty zakodowaliśmy odciskami Morgana o długości 64.



Rysunek 21. *Przedstawienie przykładowej reakcji jako charakter przemiany zapisany w notacji SMILES.*

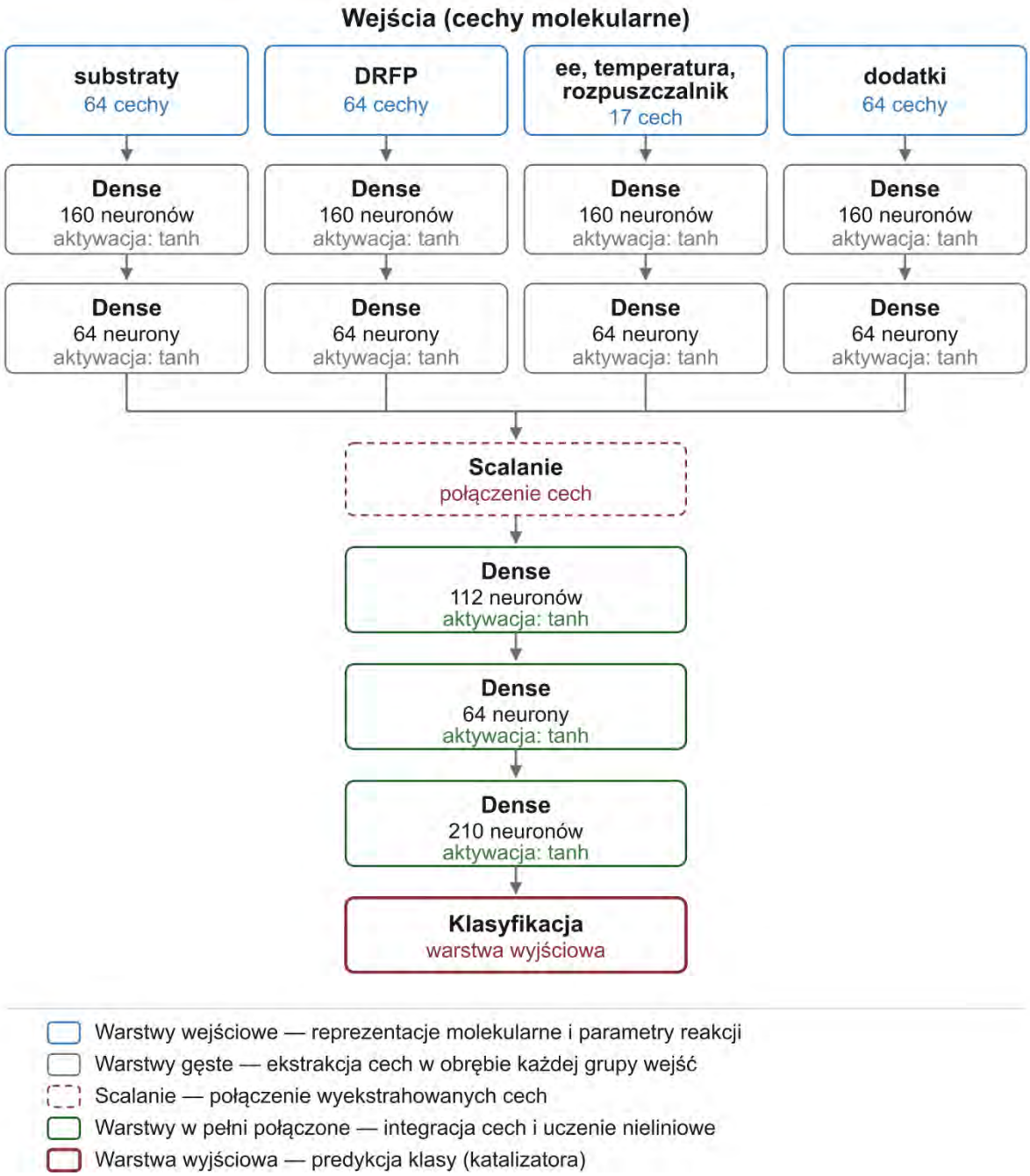
Osobnym elementem porządkowania danych było zastosowanie algorytmu Badowskiego, opracowanego w zespole prof. Grzybowskiego[71]. Algorytm wyodrębnia z zapisu reakcji centrum reakcyjne i określa typ przemiany, na przykład redukcję grupy karbonylowej (Rysunek 21) albo cykloaddycję [2+2], zwracając go w postaci zapisu SMILES obejmującego wyłącznie atomy biorące udział w przemianie. Tak uproszczony zapis opisaliśmy tymi samymi deskryptorami co pełny schemat reakcji, otrzymując wektor charakteryzujący sam charakter przemiany.

Rozpuszczalniki opisaliśmy szesnastoma właściwościami fizykochemicznymi zaczerpniętymi z tablic Stenutza[72]. Były to między innymi polaryzowalność, moment dipolowy i stała dielektryczna. Nad sposobem opisu rozpuszczalników oraz w poszukiwaniach źródeł właściwości fizykochemicznych współpracowaliśmy z dr. Kulczykowskim. Moją dodatkowo było ujednolicenie wszystkich nazw i zapisu rozpuszczalników, aby umożliwić przypisanie tych właściwości do naszych rekordów. Opracowałam też jaki rozpuszczalnik przypisać do reakcji na podstawie sprawdzania publikacji źródłowych w przypadku obecności wielu rozpuszczalników przypisanych do jednej reakcji – a była to sytuacja dosyć częsta. Dodatki zakodowano binarnymi odciskami Morgana o długości 64, przy czym ich brak reprezentowano wektorem zerowym, co zachowuje spójność wymiarową wejścia. Temperaturę i nadmiar enancjomeryczny wprowadzono bezpośrednio jako wielkości skalarne. Pierwotny wektor był dłuższy, natomiast dr Kulczykowski pracował nad jego skróceniem bez utraty zdolności predykcyjnych, aby poprawić generalizację i czytelność analizy istotności cech.

Nietypowym rozwiązaniem, które warto wyjaśnić, jest potraktowanie nadmiaru enancjomerycznego jako cechy wejściowej, a nie jako zmiennej przewidywanej. Uzasadnienie jest chemiczne: jeżeli w zbiorze występują reakcje tego samego typu prowadzone z różnymi ligandami i dające różne wartości *ee*, to model powinien nauczyć się, który ligand odpowiada wynikowi niskiemu, a który wysokiemu. Podczas predykcji ustawialiśmy następnie *ee* na stałą wartość 99%, dzięki czemu model domyślnie proponuje układy zmierzające ku wysokiej enancjoselektywności.

6.1.3 Model, jego optymalizacja i ocena

Tak zdefiniowane wektory stanowiły wejście modelu opartego na sieci neuronowej, opracowanego przez dr. Kulczykowskiego. Wybór tej architektury poprzedziła analiza porównawcza z innymi metodami, w tym z lasem losowym, które dawały wyniki nieco gorsze. Model zwraca pięć ligandów możliwych do zastosowania w zadanej reakcji i zadanych warunkach. Architekturę i hiperparametry, czyli funkcję aktywacji, rozmiary warstw, sposób inicjalizacji wag, parametry uczenia i sposób łączenia warstw, były optymalizowane biblioteką Optuna[73] w 300 iteracjach. Architekturę finalnego modelu przedstawia Rysunek 22.



***Rysunek 22**. Architektura zastosowanej sieci neuronowej.*

Ocenę przeprowadzono w schemacie pięciokrotnej walidacji krzyżowej. Obok dokładności raportowano ważoną średnią miary F1, istotną ze względu na nierównomierny rozkład klas, wynikający z bardzo różnej częstości występowania poszczególnych ligandów. Dokładność wskazania na pierwszej pozycji wyniosła średnio 0,811, a w obrębie pierwszych pięciu propozycji - 0,888.

Sprawdzono również, czy wynik zależy od rodzaju i długości zastosowanego fingerprintu molekularnego. Dr Kulczykowski przebadał wektory o długościach od 64 do 1024 w wariantach RDKit, Atom Pair, Topological Torsion, Morgan Atom Invariant oraz Morgan. Dokładność pozostawała zbliżona, oscylując wokół 0,8 niezależnie od reprezentacji. Wynik ten należy jednak

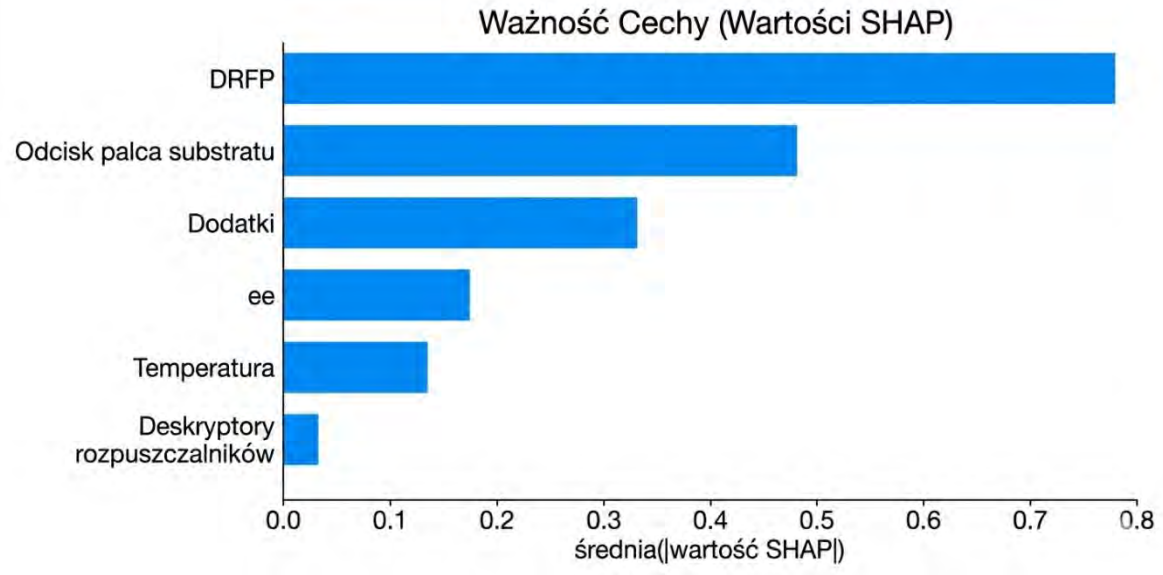
czytać ostrożnie: wszystkie porównywane warianty były odciskami molekularnymi, a porównania z deskryptorami innego typu (np. z biblioteką Mordred) w tej pracy nie prowadzono.

Osobnym pytaniem było, czy model nie faworyzuje po prostu ligandów najpopularniejszych. Jest to standardowa kontrola w uczeniu maszynowym: model użyteczny musi wypadać lepiej niż punkt odniesienia wskazujący za każdym razem najczęstszą klasę. Dr Kulczykowski i dr R. Roszak zbadali w tym celu jakość predykcji w funkcji liczby wystąpień liganda w zbiorze treningowym. Dla ligandów rzadko reprezentowanych dokładność jest niska, co odpowiada oczekiwaniom, natomiast dla ligandów występujących w co najmniej pięciu reakcjach stabilizuje się natomiast na poziomie około 80%. Dostępność danych wpływa więc na jakość predykcji w przypadkach rzadkich, ale model nie ogranicza się do odtwarzania popularności.

Pełna analiza zdolności predykcyjnych modelu wspomniana powyżej jest dostępna w materiałach uzupełniających do publikacji *Machine Learning Algorithm Guides Catalyst Choices for Magnesium-Catalyzed Asymmetric Reactions*.

6.1.4 Analiza istotności cech oraz wpływ warunków reakcji

Dysponując oceną jakości modelu, przeanalizowaliśmy wkład poszczególnych cech wejściowych metodą wartości SHAP[42], pozwalającą określić wpływ każdej zmiennej zarówno globalnie, jak i dla pojedynczej predykcji. Największe znaczenie miały deskryptory opisujące samą przemianę - odciski DRFP[74] oraz cechy strukturalne substratów kodowane odciskami Morgana. W dalszej kolejności istotne okazały się informacje o dodatkach, a następnie parametry eksperymentalne: wartość *ee*, temperatura i rozpuszczalnik (Rysunek 23).



Rysunek 23. *Istotność cech określona na podstawie wartości SHAP. Długość słupków na osi x odpowiada średnim wartościom bezwzględnym SHAP dla danej cechy, zsumowanym dla wszystkich katalizatorów.*

Rozkład ten jest chemicznie sensowny i zgodny z intuicją: o wyborze liganda decyduje przede wszystkim to, jaka przemiana ma zajść i na jakim substracie, a warunki modyfikują ten wybór, nie zastępując go. Analiza wpływu dodatków i rozpuszczalnika, przedstawiona w materiałach uzupełniających publikacji, pokazuje przy tym, że zmiana rozpuszczalnika przy zachowaniu pozostałych parametrów prowadzi do zauważalnej zmiany rankingu ligandów, natomiast wartość nadmiaru enancjomerycznego i temperatura wpływają na różnorodność proponowanych układów marginalnie. Wyniki te wskazują, że sieć nauczyła się zależności między reakcją a warunkami jej prowadzenia.

6.1.5 Walidacja eksperymentalna

Wskazania modelu sprawdziłam doświadczalnie. Wyniki poniższej walidacji eksperymentalnej opracowałam od wyboru reakcji, doboru i otrzymania substratów po końcową interpretację wyników.

Na podstawie analizy predykcji modelu wytypowałam przypadki, które uznałam za wystarczająco istotne, by poddać je weryfikacji eksperymentalnej. Przy wyborze kierowałam się dwoma kryteriami: uwzględniłam reakcje, dla których model przewidywał wyższą skuteczność niż uzyskiwana dotychczas przy użyciu katalizatorów magnezowych oraz wzięłam pod uwagę układy, w których zastosowanie katalizy magnezowej stanowiłoby alternatywę dla metod opartych na innych metalach.

Walidacja eksperymentalna została przeprowadzona w ujednoliconych warunkach - wszystkie reakcje testowano w skali 0,1 mmol substratu z użyciem 1–1,5 mL rozpuszczalnika. Takie podejście pozwoliło skrócić czas walidacji i zapewnić porównywalność wyników pomiędzy poszczególnymi reakcjami. Celem nie była pełna optymalizacja, lecz odpowiedź na pytanie, czy proponowany przez model ligand indukuje nadmiar enancjomeryczny, który można traktować jako punkt wyjścia do dalszych badań. Ustanowiłam modelowe warunki reakcji dla pracy w tym modelu, żeby zapewnić porównywalność testowanych reakcji.

Jako pierwszy przykład zbadałam asymetryczną redukcję α-ketoestrów 1a–3a, prowadzącą do produktów **1b**, **2b** i **3b** (Rysunek 24). Reakcje te zostały wcześniej opisane w literaturze z wykorzystaniem dibutylomagnezu i chiralnego kwasu fosforowego (**L-1**) w obecności katecholoboranu, prowadzone w mezytylenie [75]. Ponieważ wspomniane przykłady wchodziły w skład zbioru treningowego, model poprawnie wskazał ligand **L-1** jako najlepszy wybór (top-1) dla produktów **1b** i **3b** oraz jako drugą propozycję (top-2) dla produktu **2b**. Wynik ten nie jest zatem zaskakujący, ale zgodny z oczekiwaniami.

Z uwagi na ogrom testowanych ligandów, na tym etapie wspomagali mnie inni członkowie zespołu. Ligandy Fenga były zakupione oraz zsyntetyzowane przez dr hab. Bartosza Zambronia. Syntezą ligandów CPA używanych w publikacji zajmował się Prof. Zbigniew Pakulski. Ja przygotowałam substraty wykorzystane w walidacji eksperymentalnej oraz przeprowadziłam wszystkie przedstawione w jej ramach reakcje i analizy otrzymanych produktów.

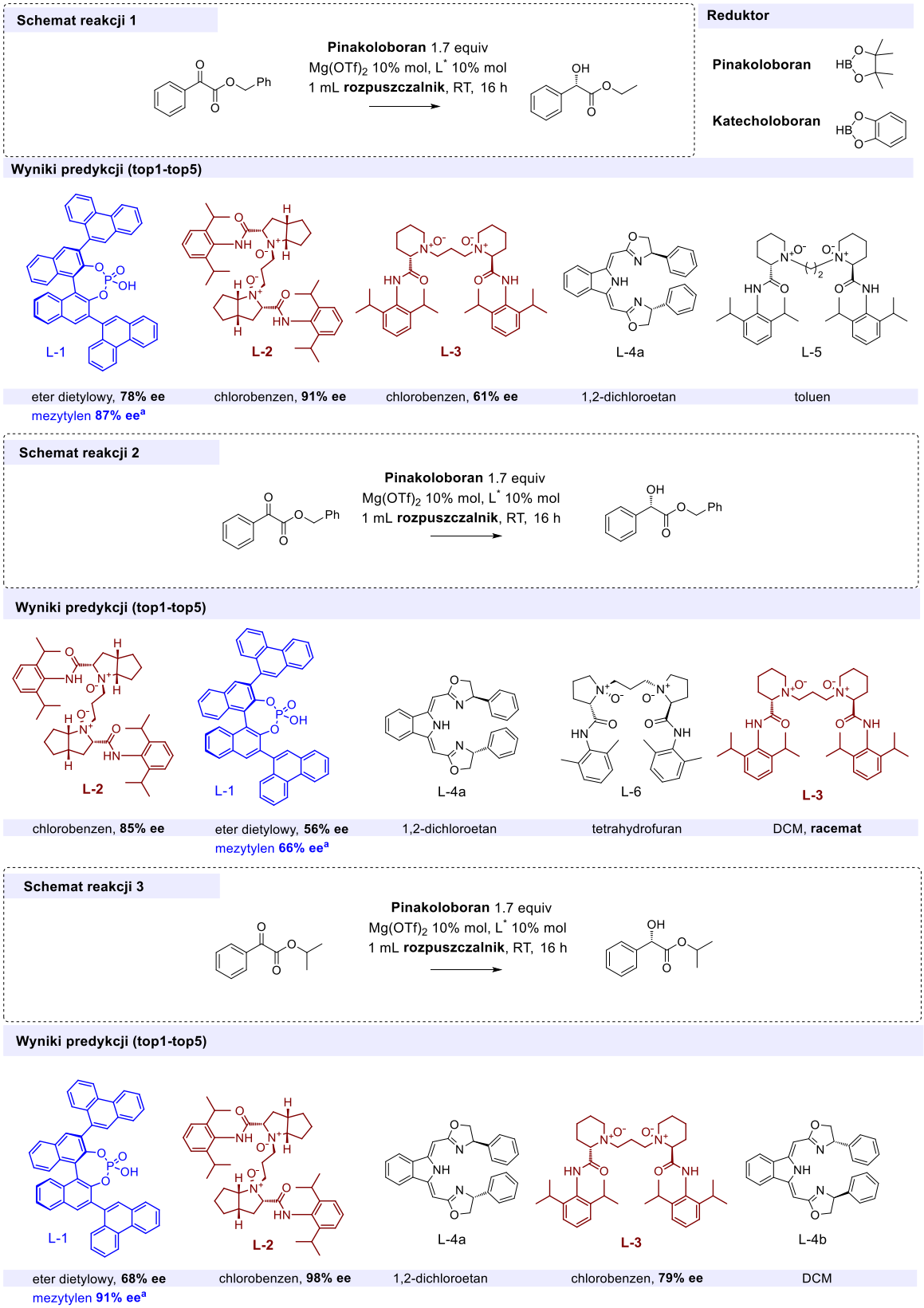
Większe znaczenie praktyczne miały jednak alternatywne propozycje ligandów generowane przez model. Pięć najwyżej ocenianych spośród nich przedstawiono na Rysunku 24. Ligandy **L-2**, **L-3**, **L-5** i **L-6** były wcześniej stosowane przez grupę Fenga [76, 77, 78] w różnych transformacjach asymetrycznych, takich jak reakcje Dielsa–Aldera i reakcje „ene", α-hydroksylowanie β-ketoestrów oraz otwarcie pierścienia azyrydyn [19]. Ligand **L-2** w połączeniu z triflanem iterbu(III) był z kolei wykorzystywany w redukcji Meerweina–Ponndorfa–Verleya α-ketoestrów [79], ponieważ jednak

zbiór danych obejmował wyłącznie katalizę magnezową, tego rodzaju przykłady nie były dostępne dla modelu podczas uczenia. Ligandy **L-4a** i **L-4b** były natomiast stosowane przez grupę Gadego [80] w enancjoselektywnej redukcji ketonów z użyciem pinakoloboranu.

Do weryfikacji eksperymentalnej wybrałam ligandy **L-2** oraz **L-3**, kierując się przede wszystkim ich stabilnością i dostępnością handlową. Szczególnie interesujące okazały się wyniki uzyskane dla układu $Mg(OTf)_2$/**L-2**, który model klasyfikował odpowiednio jako top-2, top-1 i top-2 dla produktów **1b**–**3b**. We wszystkich trzech przypadkach uzyskane wartości nadmiaru enancjomerycznego (*ee*) przewyższały wyniki osiągane z ligandem referencyjnym **L-1**: dla produktu **1b** wyniosły 91% wobec 87% (przy porównywalnej, ilościowej wydajności), dla **2b** — 85% wobec 66% (choć przy niższej wydajności: 83% wobec 97%), a dla **3b** — 98% wobec 91% (przy wydajnościach na poziomie 92–93%). Dodatkową zaletą tego układu jest możliwość prowadzenia reakcji w łagodniejszych warunkach — z użyciem stabilniejszego i wygodniejszego w obsłudze $Mg(OTf)_2$ w miejsce bardziej reaktywnego kompleksu di-*n*-butylomagnezu ($n\text{-}Bu_2Mg$).

Dla porównania zbadałam również ligand **L-3**, który model klasyfikował niżej (odpowiednio top-3, top-5 i top-4). Zgodnie z tymi przewidywaniami jego aktywność katalityczna okazała się wyraźnie niższa: dla produktu 1b uzyskano 61% *ee*, dla **2b** mieszaninę racemiczną, a dla 3b — 79% *ee*, przy wydajnościach wynoszących odpowiednio 63%, 87% i 72%.

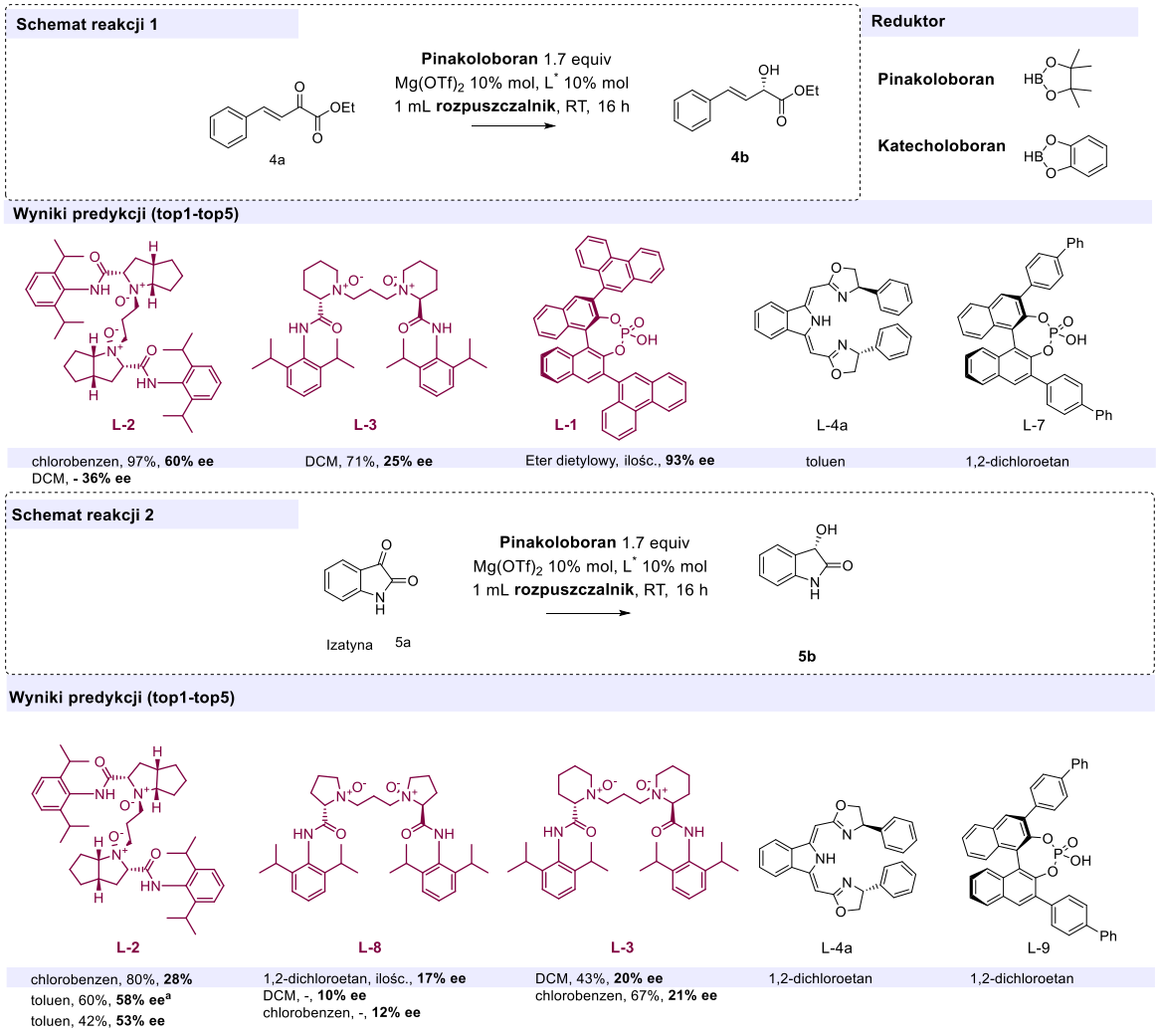
Zbadałam ponadto wpływ rozpuszczalnika sugerowanego przez model, przeprowadzając reakcje z ligandem **L-1** w eterze dietylowym zamiast mezytylenu stosowanego w pracy referencyjnej [75]. Wyniki pozostały zadowalające, jednak uzyskane wartości *ee* były niższe (odpowiednio 78% vs. 87%, 56% vs. 66% oraz 68% vs. 91%). Obserwacja ta jest spójna z analizą istotności cech (Rysunek 23), która wskazuje na relatywnie mniejszy wpływ wyboru rozpuszczalnika na przewidywania modelu.



Rysunek 24. *Predykcje modelu i wyniki walidacji eksperymentalnej dla redukcji α-ketoestrów pinakoloboranem (HBPin). Każdy wiersz przedstawia strukturę produktu wraz z ligandami sklasyfikowanymi przez algorytm na pozycjach top-1 do top-5. Wartości nadmiaru enancjomerycznego (ee), podane pod ligandami zaznaczonymi kolorem brązowym, pochodzą z eksperymentów przeprowadzonych w niniejszej pracy; konwersje we wszystkich przypadkach przekraczały 95% (szczegółowe informacje dotyczące eksperymentów znajdują się w materiałach uzupełniających). Warunki reakcji zaproponowane przez algorytm*

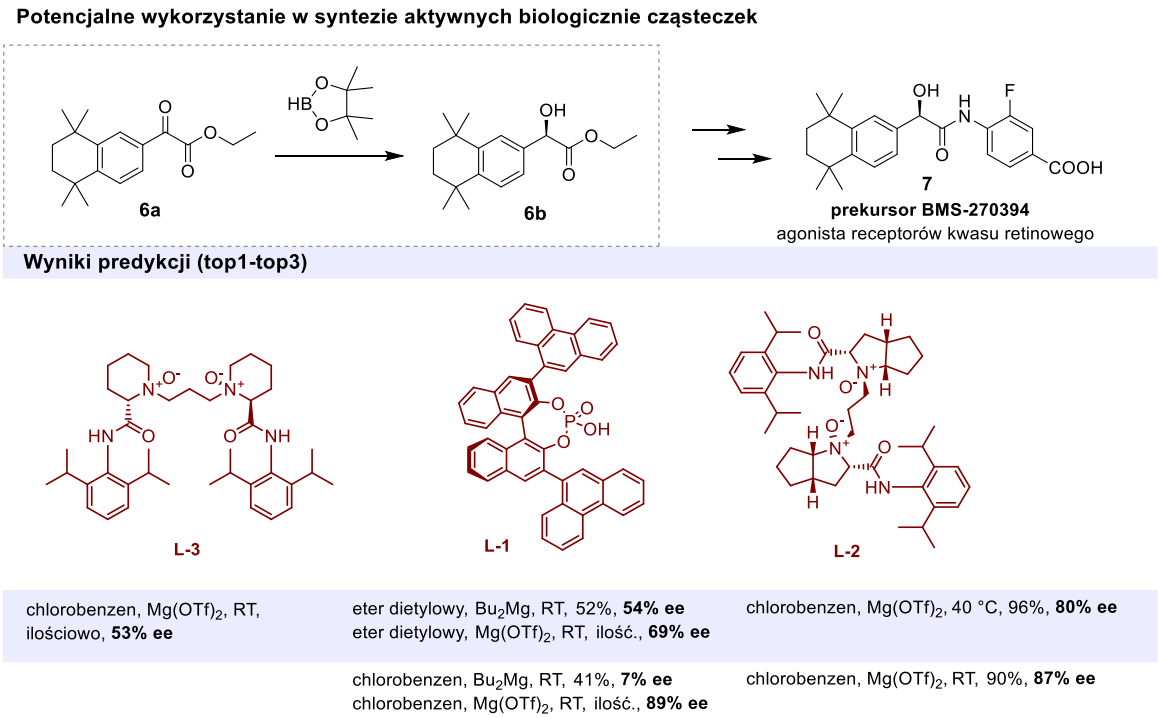
*podano na jasnoniebieskim tle. [a] Ligand **L-1** oraz odpowiadające mu wartości ee (oznaczone kolorem niebieskim) zaczerpnięto z danych literaturowych (skala 0,15 mmol, 5% **L-1**, 5% n-Bu_2Mg, katecholoboran, mezytylen jako rozpuszczalnik) [75].*

W kolejnym etapie zbadałam redukcję ketoestru **4a** zawierającego sprzężone wiązanie podwójne (górny schemat na Rysunku 25). Wybór tego substratu był podyktowany czterema względami: (1) motyw chiralnego alkoholu allilowego obecny w produkcie **4b** występuje w licznych naturalnych produktach [79]; (2) redukcja sprzężonej grupy karbonylowej zmienia charakter wiązania podwójnego z elektrofilowego na nukleofilowy, otwierając tym samym interesujące możliwości dalszych przekształceń; (3) substraty tego typu stanowią wyzwanie dla asymetrycznej redukcji ze względu na możliwość jednoczesnej redukcji wiązania podwójnego [81]; (4) reakcje tego rodzaju są z reguły realizowane z użyciem kompleksów rutenu [82].



***Rysunek 25.** Predykcje modelu i wyniki walidacji eksperymentalnej dla katalizowanej magnezem redukcji β,γ-nienasyconego ketoestru oraz izatyny. Wartości wydajności izolowanej (%) oraz nadmiaru enancjomerycznego (ee) podane czarną czcionką pochodzą z eksperymentów przeprowadzonych w niniejszej pracy (szczegółowe informacje eksperymentalne znajdują się w materiałach uzupełniających). W trzech przypadkach o niskiej indukcji asymetrycznej (**4b**-**L2**-DCM; **5b**-**L8**-DCM; **5b**-**L8**-chlorobenzen) produktów nie izolowano, a zatem wydajności nie zostały wyznaczone. Warunki reakcji na jasnoniebieskim tle (w tym rozpuszczalnik i temperatura) zostały zaproponowane przez algorytm; warunki na białym tle pochodzą z*

dodatkowych eksperymentów przeprowadzonych przeze mnie w oparciu o chemiczną intuicję. [a] Reakcja z katecholoboranem zamiast HBPin.



Rysunek 26. *Wyniki predykcji i walidacji eksperymentalnej przeprowadzone dla prekursora biologicznie aktywnej cząsteczki BMS-270394.*

W literaturze z lat osiemdziesiątych XX wieku opisano dwie metody zastosowania magnezu w asymetrycznej redukcji substratu **4a** [83, 84]. Należy jednak zaznaczyć, że żadna z nich nie ma charakteru katalitycznego, ponieważ obie wymagają użycia 1,1 równoważnika chiralnego liganda będącego mimetykiem NADH oraz 1 równoważnika soli magnezowej.

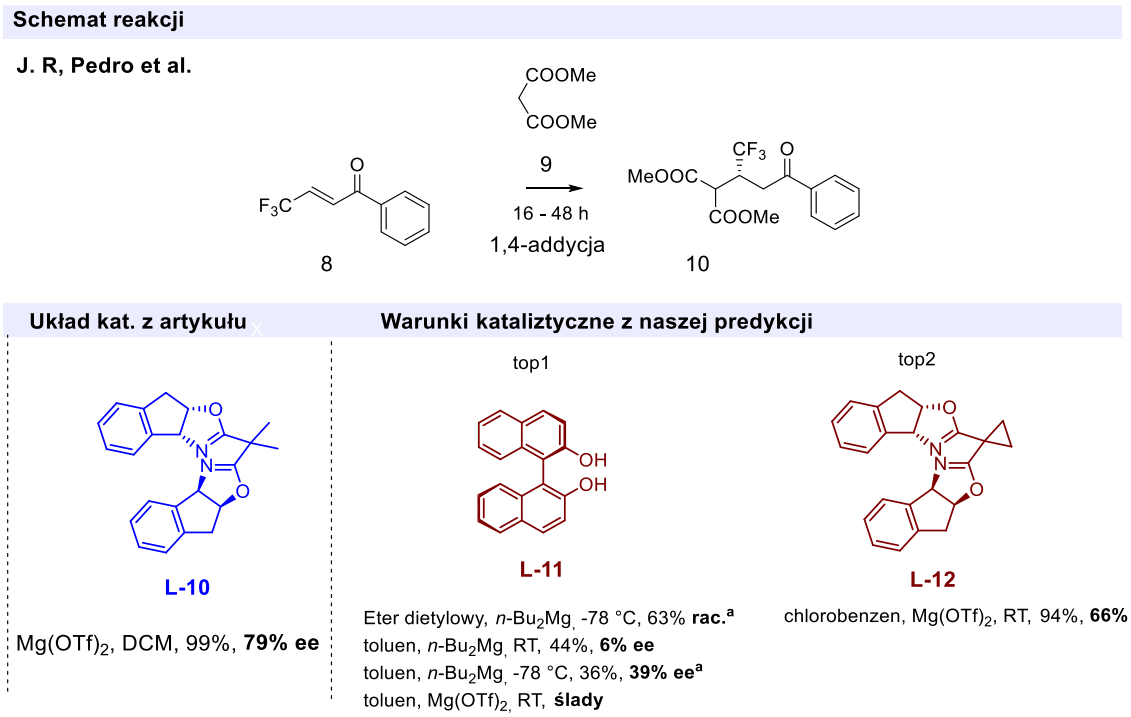
Wyniki uzyskane z ligandami **L-2** (predykcja top-1) i **L-3** (predykcja top-2, dostępny handlowo), które wynoszą odpowiednio 60% i 25% *ee* przy dobrej wydajności, na tym tle można uznać za obiecujący punkt wyjścia do dalszej optymalizacji. Wyniki te są tym bardziej zachęcające, że dotyczą układu, dla którego nie istnieją wcześniejsze przykłady katalizy magnezowej, co czyni ten problem szczególnie trudnym. Ligand sklasyfikowany przez model na pozycji top-3 był wprawdzie stosowany wcześniej w redukcji ketoestrów, jednak wyłącznie dla aromatycznych ketoestrów pozbawionych łańcucha alkilowego [75]. Nieoczekiwanie zastosowanie liganda **L-1** w eterze dietylowym, który nie jest typowo kojarzony z tą chemią, dało produkt **4b** z ilościową wydajnością i 93% nadmiarem enancjomerycznym.

Kolejnym zbadanym przeze mnie przykładem była redukcja izatyny, oznaczonej jako **5a** na Rysunku 25. Pochodne izatyny z podstawnikiem przy atomie azotu były wcześniej stosowane w asymetrycznych reakcjach Dielsa–Aldera oraz reakcjach „ene" katalizowanych kompleksami magnezu z ligandami Fenga[19]. Model wskazał, że te same ligandy (**L-2**, **L-8**, **L-3**) powinny być obiecujące (top-3) również w asymetrycznej redukcji α-ketoamidów. Wynik ten uznałam za szczególnie interesujący, ponieważ chiralne produkty typu 2-aminoalkoholu (np. **5b**) stanowią istotny motyw strukturalny w projektowaniu biologicznie aktywnych związków [85].

W eksperymencie wszystkie testowane ligandy indukowały pewien stopień asymetrii, jednak jedynie dla predykcji top-1 (**L-2**) uzyskano zadowalające wartości *ee*: 50% po zastosowaniu toluenu jako rozpuszczalnika i 58% po zamianie pinakoloboranu na katecholoboran jako czynnik redukujący. W przypadku ligandów **L-8** i **L-3**, dających niższe wartości *ee*, próba poprawy wyników przez zmianę rozpuszczalnika nie przyniosła istotnego efektu.

Chcąc wyjść poza substytucję modelowych β,γ-nienasyconych substratów, skupiłam uwagę na ketoestrze **6a**, którego asymetryczna redukcja prowadzi do **6b**, który jest prekursorem do otrzymania BMS-270394 **(7)** Rysunek 26. Przedstawiona cząsteczka, opisana w licznych patentach i publikacjach, jest agonistą receptorów kwasu retinowego o udokumentowanej aktywności przeciwnowotworowej [86, 87]. W 2002 roku firma BMS (Bristol-Myers Squibb) udoskonaliła własną syntezę, przeprowadzając asymetryczną redukcję odpowiedniego kwasu z użyciem borowodorku sodu i kwasu L-winowego, uzyskując wydajność 75% i 68% *ee*. Czystość enancjomeryczną poprawiono następnie do >99% przez ko-krystalizację z (1*R*,2*R*)-(−)-*N*,*N*'-dimetylocykloheksano-1,2-diaminą [88].

W tym kontekście model zaproponował ligand Fenga **L-3** jako top-1, **L-1** (CPA) jako top-2 oraz **L-2** jako top-3. We wszystkich przypadkach uzyskałam pozytywne wyniki. Najwyższe *ee* wyniosło 80% dla **L-2**. Optymalizacja rozpuszczalnika pozwoliła podnieść *ee* do 89% dla **L-1**, natomiast obniżenie temperatury do 87% dla **L-2**. Wydajności reakcji z $Mg(OTf)_2$ mieściły się w zakresie 90–99%, co czyni tę metodę atrakcyjną alternatywą dla procedury opisanej wcześniej[88]. Istotną rolę odegrało źródło magnezu: przykładowo w chlorobenzenie ligand **L-1** dawał zaledwie 7% *ee* przy zastosowaniu *n*-Bu_2Mg, podczas gdy z $Mg(OTf)_2$ uzyskano 89% *ee*. Wydajności z dibutylomagnezu były również niższe, oscylując na poziomie około 50%. Co ciekawe, różnica ta była znacznie mniej wyraźna w eterze dietylowym (54% vs. 69%), co podkreśla wrażliwość reakcji katalizowanych magnezem na warunki prowadzenia, która pozostaje jednym z głównych wyzwań w tej dziedzinie, a zarazem ograniczeniem modeli predykcyjnych takich jak opisywany.



Rysunek 27. *Katalizowana kompleksem magnezu reakcja addycji 1,4 — porównanie skuteczności ligandów. Ligand zaznaczony kolorem niebieskim pochodzi z najnowszej literatury; ligandy przetestowane przeze mnie oznaczono kolorem brązowym. Wartości wydajności izolowanej (%) oraz nadmiaru enancjomerycznego (ee) podane czarną czcionką pochodzą z eksperymentów przeprowadzonych w niniejszej pracy (szczegółowe*

informacje eksperymentalne znajdują się w materiałach uzupełniających). Warunki reakcji na jasnoniebieskim tle (rozpuszczalnik i temperatura) zostały zaproponowane przez algorytm; warunki na białym tle pochodzą z dodatkowych eksperymentów przeprowadzonych przeze mnie w oparciu o chemiczną intuicję. [a] Reakcje mieszano w temperaturze −78 °C przez 24 godziny, a następnie ogrzewano do temperatury pokojowej.

Ostatnia seria walidacji dotyczyła reakcji katalizowanych magnezem opublikowanych już po skompletowaniu zbioru danych, a zatem nieznanych modelowi w trakcie uczenia. Jednym z takich przykładów była addycja Michaela opisana przez grupę Pedra [89], zilustrowana na Rysunku 27. Model nie miał dostępu do informacji o ligandzie **L-10** zastosowanym w tej pracy, zaproponował jednak jego bliski analog strukturalny (**L-12**) jako wybór top-2, natomiast na pozycji top-1 wskazał popularny i dostępny handlowo BINOL (**L-11**). Oba ligandy przetestowałam zarówno w warunkach sugerowanych przez model, jak i w kilku chemicznie intuicyjnych wariantach alternatywnych, np. w niekoordynującym toluenie zamiast eterze dietylowym, w którym **L-11** z $Mg(OTf)_2$ tworzy zawiesinę. Pozytywny wynik dla **L-12** (66% *ee*) można było w pewnym stopniu przewidzieć na podstawie jego bliskiego podobieństwa strukturalnego do **L-10**. Natomiast uzyskanie 39% *ee* dla stosunkowo prostego liganda BINOL było dla mnie zaskoczeniem i sugeruje, że dalsza optymalizacja warunków może okazać się owocna, co jest zjawiskiem często obserwowanym w przypadku ligandów z grupy BINOL [90, 91]. Analizując ten przypadek szczegółowo, stwierdziłam, że zbiór treningowy zawierał cztery typy reakcji katalizowanych BINOL-em: addycję Mannicha, hetero reakcję Dielsa–Aldera, otwarcie pierścienia azyrydyny oraz addycję Michaela do nienasyconych ketoestrów. Sugestia modelu najprawdopodobniej odzwierciedlała przykłady z tych ostatnich, ze względu na podobieństwo strukturalne substratów [92, 93].

6.1.6 Podsumowanie

Opracowany model uczenia maszynowego identyfikował zależności pomiędzy strukturą substratów, typem reakcji, warunkami prowadzenia oraz zastosowanym ligandem chiralnym. Jego przydatność sprawdzono eksperymentalnie w redukcjach α-ketoestrów, α,β-nienasyconych ketoestrów i α-ketoamidów oraz w addycji Michaela. W trzech redukcjach α-ketoestrów rekomendowany układ $Mg(OTf)_2$/**L-2** dawał wyższy nadmiar enancjomeryczny niż układ literaturowy. Dla nienasyconego ketoestru i izatyny uzyskano natomiast przykłady reakcji katalizowanych magnezem, których wcześniej nie opisano dla tych klas substratów. W części przypadków możliwe było również zastąpienie labilnego di-*n*-butylomagnezu trwalszym i tańszym $Mg(OTf)_2$.

Model lepiej sprawdzał się w predykcji układu katalitycznego niż w sugerowaniu warunków reakcji. Wynika to między innymi ze złożonego wpływu rozpuszczalnika i warunków na katalizę magnezową oraz z nierównomiernego rozkładu danych literaturowych. Ważniejszym ograniczeniem jest to, że model zawsze wskazuje najlepszy spośród dostępnych katalizatorów, również wtedy, gdy dana transformacja może w ogóle nie zachodzić w obecności magnezu. Predykcję należy więc rozumieć warunkowo, jako odpowiedź na pytanie, który katalizator warto wybrać, jeżeli dana reakcja jest możliwa do przeprowadzenia z udziałem magnezu.

Dalszy rozwój takiego podejścia wymaga przede wszystkim lepszych zbiorów danych, obejmujących również reakcje nieudane i układy katalizowane różnymi metalami. Wyniki tej pracy pokazały jednocześnie, że sama liczba rekordów nie jest wystarczająca. Dane pozyskane automatycznie wymagają kontroli chemicznej, ujednolicenia i ręcznej weryfikacji. Dopiero na tak przygotowanych zbiorach zasadne jest porównywanie bardziej zaawansowanych modeli i

reprezentacji uwzględniających między innymi geometrię 3D lub właściwości elektronowe cząsteczek.

Moją rolą było zaprojektowanie i przygotowanie zbioru danych, określenie jego zawartości oraz ręczna weryfikacja i ujednolicenie informacji chemicznych. Odpowiadałam również za stronę chemiczną procesu budowy modelu, ocenę poprawności kolejnych rozwiązań i proponowanie zmian w sposobie jego konstrukcji i interpretacji. Zaplanowałam walidację eksperymentalną, przygotowałam wymagane substraty oraz przeprowadziłam wszystkie przedstawione reakcje i analizy otrzymanych produktów. Do mojego wkładu należała także interpretacja wyników, przygotowanie grafik, współtworzenie aplikacji od strony merytorycznej oraz współautorstwo publikacji. Całość pracy pokazała, że dane literaturowe mogą zostać wykorzystane nie tylko do opisu znanych zależności, ale również do praktycznego wskazywania punktów wyjścia do eksperymentalnej optymalizacji nowych reakcji asymetrycznych.

Badania opisane w tym rozdziale zostały opublikowane w pracy: P. Baczewska, M. Kulczykowski, B. Zambroń, J. A. Jaszczewska-Adamczak, Z. Pakulski, R. Roszak, B. A. Grzybowski, J. Mlynarski, Machine Learning Algorithm Guides Catalyst Choices for Magnesium-Catalyzed Asymmetric Reactions, *Angew. Chem. Int. Ed.* **2024**, 63, e202318487.

6.2 Przewidywanie reaktywności i enancjoselektywności na podstawie małych zbiorów danych eksperymentalnych na przykładzie reakcji katalizowanych kompleksami magnezu

Model opisany w rozdziale 6.1 korzystał z danych literaturowych i wskazywał dla zadanej reakcji najbardziej obiecujące spośród układów już znanych, dając punkt wyjścia do dalszej optymalizacji. Jego ograniczeniem była możliwość przewidywania wyłącznie w obrębie chemii reprezentowanej w zgromadzonych danych oraz ograniczona skuteczność w doborze warunków. W badaniach opisanych poniżej odwróciliśmy ten problem: zamiast dużego i niejednorodnego zbioru literaturowego wykorzystaliśmy zbiory niewielkie, lecz wewnętrznie spójne, pochodzące z naszego laboratorium.

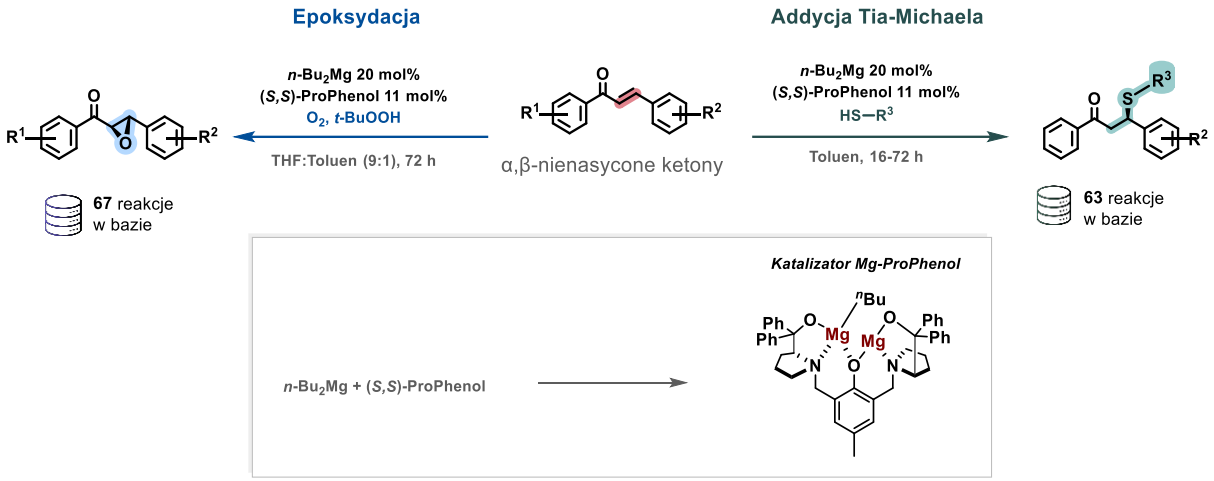
Rozwój nowej reakcji asymetrycznej rzadko prowadzi do setek systematycznie przebadanych kombinacji. Znacznie częściej powstaje niewielki zakres struktur otrzymanych w jednolitych warunkach, obejmujący zarówno reakcje udane, jak i substraty reagujące słabo albo wcale. Z punktu widzenia uczenia maszynowego zbiór taki jest mały, ma jednak zaletę, której nie ma żaden zbiór literaturowy: wszystkie wyniki powstały według tego samego protokołu, a ich pochodzenie i jakość są znane. Co więcej, reakcje nieudane nie są tu brakiem informacji, lecz pełnoprawnym wynikiem wyznaczającym granicę reaktywności układu. Postawiłam zatem pytanie, ile informacji da się wydobyć ze zbioru, który powstaje jako produkt uboczny zwykłej pracy laboratorium chemii organicznej, a nie jako materiał zaprojektowany pod modelowanie.

Celem nie było wykonanie tysięcy dodatkowych doświadczeń na potrzeby modelu, lecz wsparcie dwóch typowych zadań chemika: oceny, czy nowy substrat w ogóle przereaguje, oraz przewidzenia, jakiej enancjoselektywności można się spodziewać w obrębie badanego obszaru strukturalnego. Prospektywne wskazania mają wtedy służyć bardziej racjonalnemu wyborowi kolejnych substratów i rozszerzaniu zakresu metody.

6.2.1 Układ Mg–ProPhenol i reakcje modelowe

Ligand ProPhenol należy do ligandów bifunkcyjnych: łączy fenolowe centrum wiążące metal z dwoma pierścieniami pirolidynowymi wprowadzającymi chiralność i zdolnymi do działania zasadowego. W obecności n-Bu_2Mg powstaje z niego dwurdzeniowy kompleks magnezu, w którym jedno centrum metaliczne aktywuje elektrofil, a drugie uczestniczy w wiązaniu i aktywacji nukleofila. Układ ten stanowi podstawę obu reakcji modelowych opisanych w rozdziale 6.2.

Jako układy modelowe wybraliśmy dwie reakcje katalizowane kompleksami magnezu, opracowane wcześniej w naszym laboratorium przez dr J. Jaszczewską-Adamczak: enancjoselektywną epoksydację chalkonów oraz addycję tio-Michaela do α,β-nienasyconych ketonów katalizowaną kompleksem dwurdzeniowym (Rysunek 28). W projekcie dotyczącym addycji tio-Michaela współpracowałam przy optymalizacji reakcji, zsyntetyzowałam jeden z ligandów PropPhenolowych oraz przeprowadziłam obliczenia kwantowochemiczne, na podstawie których opracowałam model stereochemiczny tej przemiany. Dane te zaczerpnięto z naszych wcześniejszych publikacji i wykorzystano w niniejszym projekcie.[94, 95]



Rysunek 28. *Badane układy reakcyjne wykorzystane do budowy małych zbiorów danych eksperymentalnych: enancjoselektywna epoksydacja chalkonów oraz addycja tio-Michaela do α,β-nienasyconych ketonów katalizowana dwurcentrowym kompleksem magnezu.*

Epoksydację chalkonów i pokrewnych enonów prowadzi się wobec *n*-Bu₂Mg i (*S,S*)-ProPhenolu z wodoronadtlenkiem *tert*-butylu jako utleniaczem, w mieszaninie tetrahydrofuranu i toluenu.[94] Addycję tio-Michaela tioli do α,β-nienasyconych ketonów — w analogicznym układzie katalitycznym w toluenie.[95] Wybór tej pary nie był przypadkowy. Epoksydacja obejmuje stosunkowo jednorodną grupę substratów i wąski zakres selektywności, skupiony powyżej 90% *ee*, gdzie z punktu widzenia syntezy jest to zaleta, a z punktu widzenia modelowania utrudnienie, ponieważ zmienna odpowiedzi ma niewielką zmienność do wyjaśnienia. W addycji tio-Michaela zmieniane są natomiast niezależnie oba substraty w reakcji, enon i tiol, dzięki czemu przestrzeń chemiczna jest szersza, a rozkład wartości *ee* wyraźnie bardziej zróżnicowany. Zbiór epoksydacyji zawiera przy tym istotną informację o substratach niereaktywnych, a zbiór tio-Michaela dysponuje dodatkowym kontekstem mechanistycznym.

Kontekst mechanistyczny pochodził z badań nad reakcją tio-Michaela, której dane wykorzystujemy w tym projekcie. W ramach tej pracy prowadziłam i interpretowałam obliczenia DFT, które wraz z eksperymentami ^{1}H NMR wskazały na katalitycznie istotny kompleks magnezu zawierający jednocześnie tiolan i skoordynowany chalkon. Zaproponowany mechanizm zakładał więc reakcję obu substratów związanych z tym samym kompleksem, co później pozwoliło odnosić cechy istotne dla predykcji do konkretnego obrazu chemicznego układu.

6.2.2 Budowa zbiorów danych

Ze względu na znaczenie jakości danych wykorzystałam wyłącznie wyniki uzyskane w naszym laboratorium, w dobrze zdefiniowanych i porównywalnych warunkach. Pozwoliło to ograniczyć źródła zmienności typowe dla danych literaturowych, takie jak różnice w prowadzeniu reakcji, oznaczaniu nadmiaru enancjomerycznego i sposobie raportowania wyników. Wszystkie struktury i wyniki mogłam również zweryfikować bezpośrednio na podstawie dokumentacji eksperymentalnej oraz w rozmowie z dr J. Jaszczewską-Adamczak.

Oba zbiory uzupełniłam o reakcje wykonane przez dr J. Jaszczewską-Adamczak, które nie zostały uwzględnione w jej publikacjach dotyczących epoksydacji i addycji tio-Michaela. Przygotowałam również pełne zapisy reakcji, zwracając szczególną uwagę na spójne przypisanie

konfiguracji centrów stereogenicznych w produktach na podstawie związków wcześniej scharakteryzowanych.

Tak przygotowane dane stanowią wymagający przypadek dla regresji. Jeżeli większość przykładów odpowiada wysokiej enancjoselektywności, model może uzyskiwać pozornie dobre wyniki przez przewidywanie wartości zbliżonych do najczęściej występujących, bez rzeczywistego uchwycenia zależności pomiędzy strukturą a selektywnością. Jednocześnie obecność licznej grupy substratów niereaktywnych umożliwiła sformułowanie osobnego zadania klasyfikacji reaktywności. Rozdzielenie tych dwóch problemów było istotne również z chemicznego punktu widzenia, ponieważ dla reakcji, w której produkt nie powstaje, wartość *ee* nie jest zdefiniowana i nie może być traktowana jako 0%.

Jednorodność zbioru epoksydacji była widoczna nie tylko w rozkładzie enancjoselektywności, ale również w podobieństwie struktur. Średnie podobieństwo strukturalne wynosiło 0,878, a dla najbardziej zbliżonych odrębnych przykładów sięgało 0,983. Oznacza to, że wiele substratów zajmuje bardzo podobny obszar przestrzeni chemicznej, co ogranicza ilość informacji dostępnej dla modelu i zwiększa ryzyko, że będzie on przede wszystkim odtwarzał wartość typową dla tego zbioru.

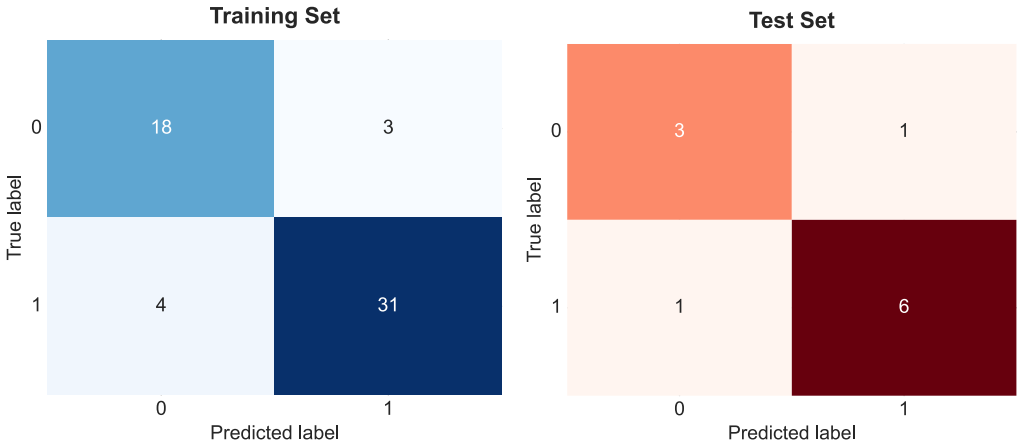
6.2.3 Rozdzielenie zadania predykcyjnego

Pierwsze modele regresyjne dla epoksydacji działały słabo, a analiza przyczyn tego niepowodzenia okazała się jednym z najważniejszych etapów pracy. Rozkład danych wskazywał na dwa odrębne problemy. Z jednej strony zbiór zawierał liczne substraty niereaktywne, a z drugiej większość reakcji zachodzących prowadziła do wysokich wartości *ee*. Uzgodniliśmy więc rozdzielenie zadania na klasyfikację reaktywności oraz regresję enancjoselektywności prowadzoną wyłącznie dla reakcji, w których produkt powstawał i można było oznaczyć *ee* (Rysunek 29, 30).

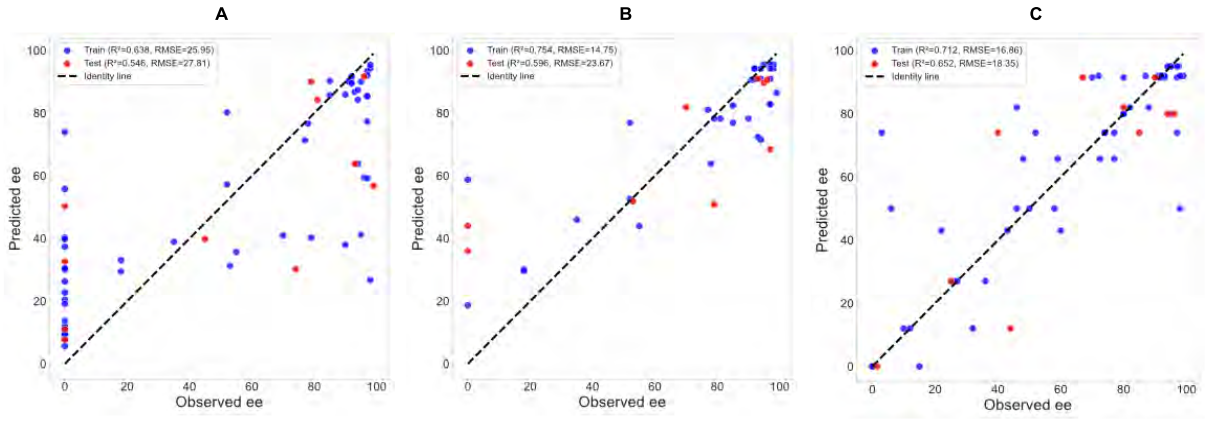
Takie rozdzielenie wynika przede wszystkim z chemicznego znaczenia danych. Brak produktu nie oznacza 0% *ee*, ponieważ dla reakcji, która nie zachodzi, enancjoselektywność nie jest zdefiniowana. Za niereaktywne uznawałam więc wyłącznie reakcje, w których nie obserwowano powstawania produktu. Wartość 0% *ee* nie była kryterium braku reaktywności. Odpowiadałam za odpowiednie przygotowanie danych, dlatego reakcje nieudane wykorzystano wyłącznie w klasyfikacji, natomiast modele regresyjne trenowano na podzbiorze reakcji zachodzących.

Samo połączenie klasyfikacji i regresji nie jest nową koncepcją uczenia maszynowego, czego przykład omówiono w podrozdziale 4.5.1. W tej pracy istotne było zastosowanie takiego podejścia do małego, realistycznego zbioru z katalizy asymetrycznej, w którym nieudane reakcje stanowią ważną informację o zakresie reaktywności substratów. Rozdzielenie obu problemów zdecydowało również o tym, czy z dostępnych danych można uzyskać użyteczny model.

Klasyfikator dawał stabilne wyniki w 120 badanych konfiguracjach. Najlepszy był model gradient boosting[96] wykorzystujący sześć deskryptorów wyprowadzonych z produktu, obejmujących parametry Sterimol[33, 97, 98] i zliczenia fragmentów CircuS[14, 30]. Dla zbioru 67 przykładów, obejmującego 56 reakcji treningowych i 11 testowych, zbalansowana dokładność wyniosła 0,87 dla zbioru treningowego i 0,80 dla testowego, a w walidacji krzyżowej 0,74 ± 0,08. Za implementację i prowadzenie prac związanych z modelowaniem odpowiadał Damian Nowak, z którym współpracowałam przy doborze założeń i interpretacji wyników.



***Rysunek 29**. Macierze pomyłek najlepszego klasyfikatora reaktywności dla epoksydacji chalkonów, wyznaczone dla zbioru treningowego i testowego.*



***Rysunek 30.** Przewidywane i obserwowane wartości nadmiaru enancjomerycznego (ee) dla modeli regresyjnych. (A) Epoksydacja chalkonów dla pełnego zbioru danych. (B) Epoksydacja chalkonów po wykluczeniu substratów niereaktywnych. (C) Reakcja tio-Michaela. Jakość modeli poprawiała się po ograniczeniu zbioru do substratów reaktywnych, a najbardziej stabilne wyniki uzyskano dla bardziej zróżnicowanego strukturalnie zbioru reakcji tio-Michaela, dla którego R^2 w zbiorze testowym osiągało wartość do 0,65.*

6.2.4 Reprezentacja chemiczna i dobór modeli

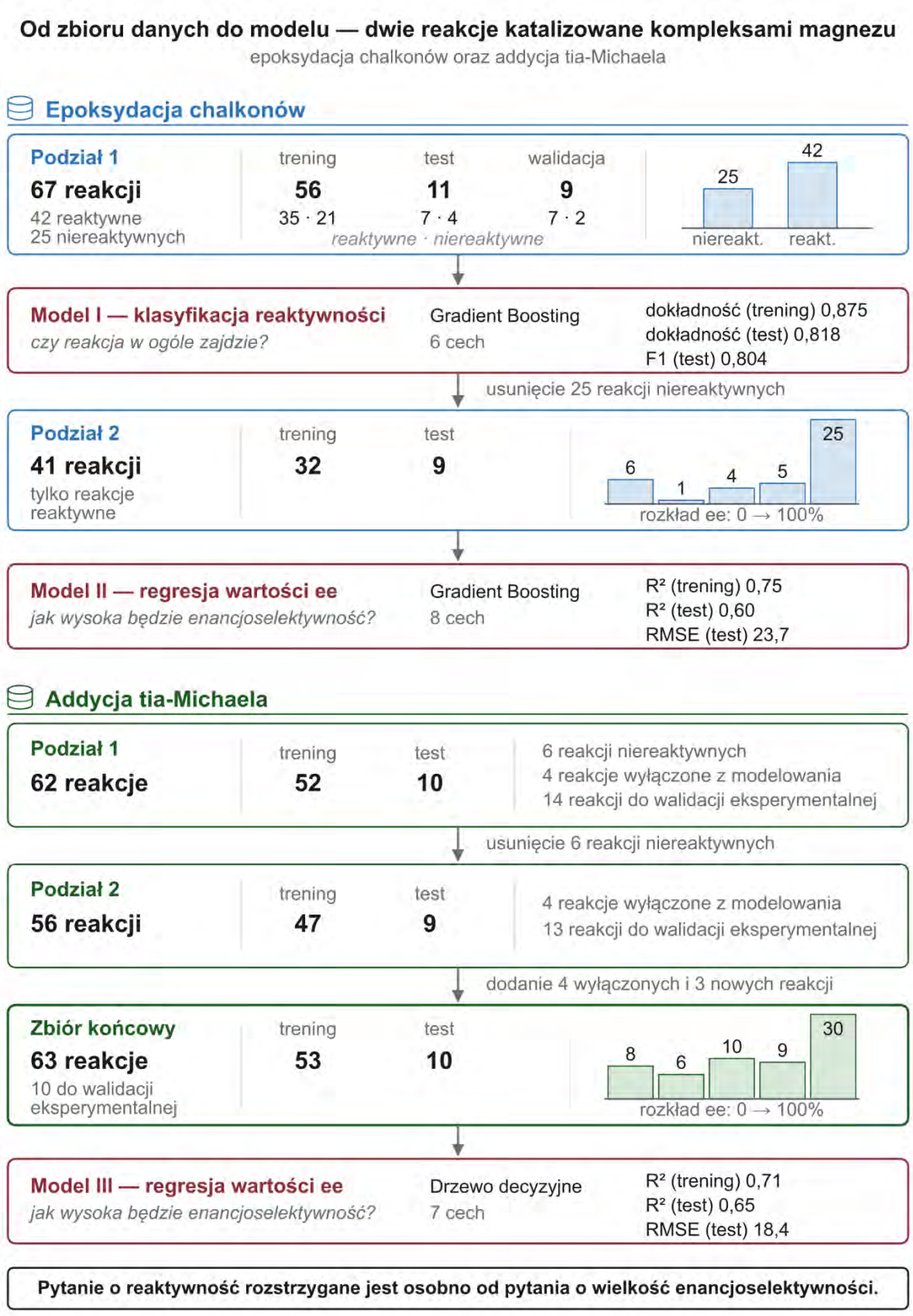
Stronę obliczeniową prowadzili współautorzy z Wydziału Chemii Uniwersytetu im. Adama Mickiewicza w Poznaniu. Generowanie deskryptorów, redukcję liczby cech, optymalizację hiperparametrów oraz przygotowanie metryk i wykresów wykonywał mgr Damian Nowak. Moją rolą była kontrola merytoryczna tej części od strony chemicznej, ocena kolejnych modeli i proponowanie zmian wtedy, gdy rozwiązania formalnie poprawne z punktu widzenia uczenia maszynowego nie dawały użytecznych odpowiedzi chemicznych.

Pierwsze modele regresyjne opierały się na niewielkiej liczbie cech wybranych z biblioteki Mordred[31] i działały słabo. Taki opis mógł być zbyt uproszczony, aby uchwycić subtelne różnice strukturalne decydujące o enancjoselektywności badanych reakcji. Z tego względu zaczęliśmy

szukać bogatszych reprezentacji, które lepiej opisywałyby zarówno budowę przestrzenną substratów, jak i ich lokalne właściwości strukturalne.

Przetestowano szeroki zakres reprezentacji obejmujących (oprócz deskryptorów Mordred[31]) fingerprinty Morgana[28, 99], osadzenia ChemBERTa[100], zliczenia fragmentów CircuS i ChyLine[14, 30], parametry Sterimol[33, 97, 98] oraz proste deskryptory elektronowe oparte na ładunkach Gasteigera[101]. Przy tak małych zbiorach wiele konfiguracji łatwo prowadziło do przeuczenia, dlatego dobór sposobu opisu cząsteczek miał duże znaczenie. Zaproponowałam wykorzystanie deskryptorów CircuS i ChyLine, natomiast dr R. A. Bachorz zaproponował zastosowanie modelu ChemBERTa. Ostatecznie najlepsze wyniki uzyskano dla reprezentacji opartych na Sterimol i CircuS. Osadzenia ChemBERTa zostały przetestowane, ale nie weszły do żadnego z modeli końcowych, a zestawy bez nich dawały lepsze wyniki w obu zadaniach regresyjnych.

W trakcie prac zmieniliśmy również początkowe kryterium wyboru modeli, które premiowało możliwie małą liczbę cech. W praktyce zbyt silne uproszczenie prowadziło do utraty istotnych różnic pomiędzy substratami i ograniczało wartość chemiczną modelu. Współpracowaliśmy więc w kolejnych etapach modyfikacji modeli, proponując rozwiązania, które pozwalały zachować istotną informację strukturalną, a jednocześnie ograniczać przeuczenie i poprawiać zdolność generalizacji. Finalne zestawy danych i metryki modeli przedstawia Rysunek 31.



Rysunek 31. *Budowa zbiorów danych i modeli dla epoksydacji chalkonów oraz addycji tio-Michaela.*

6.2.5 Wybór zmiennej odpowiedzi

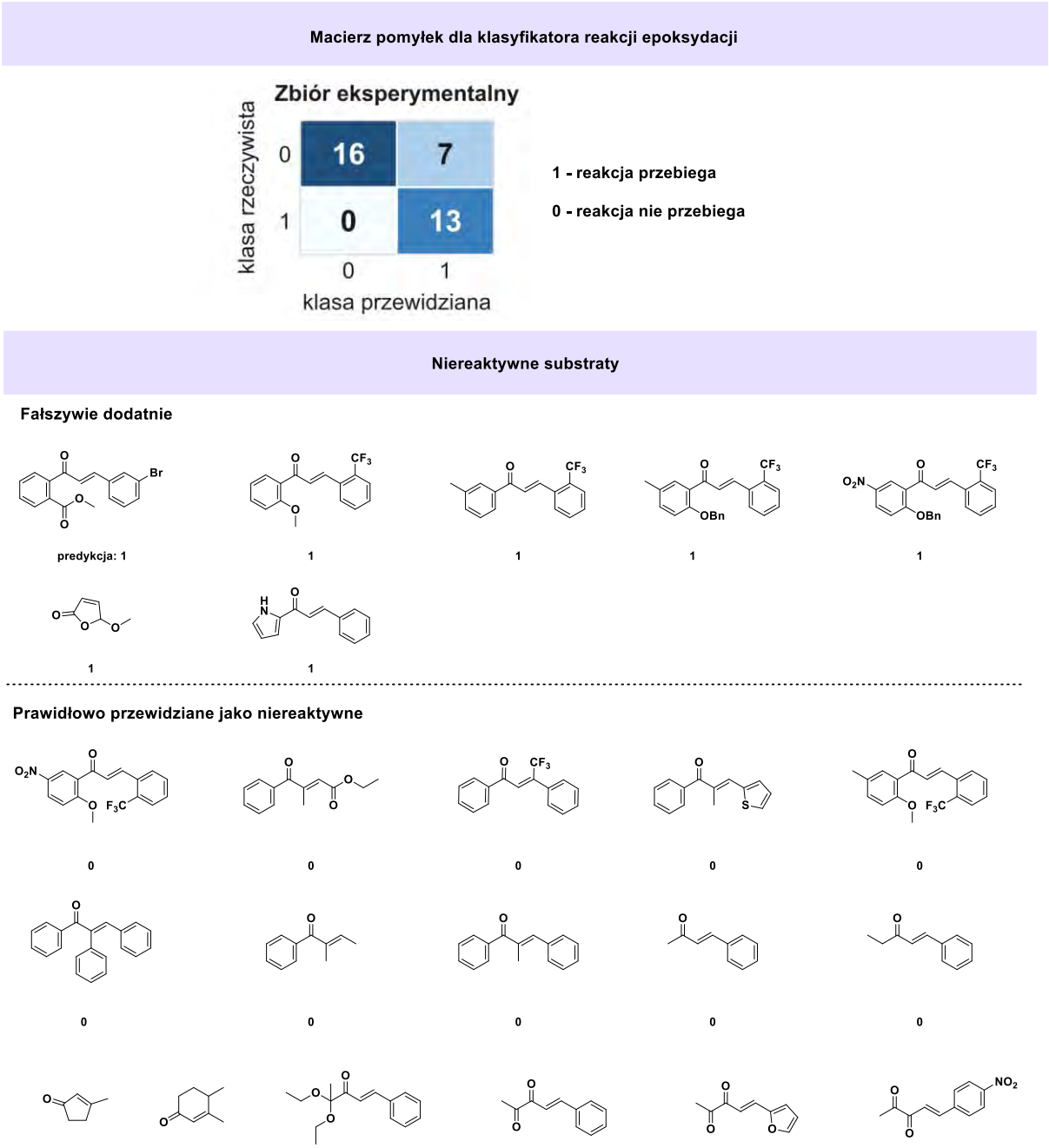
W odpowiedzi na uwagę recenzenta rozważaliśmy przeliczenie wartości *ee* na różnicę energii swobodnej aktywacji (ΔΔG‡). Po przeanalizowaniu danych uznałam jednak, że dla tego zbioru rozwiązanie to nie jest ‘jedyną poprawną odpowiedzią’. Procedura epoksydacji obejmuje chłodzenie, natlenianie, następnie ogrzewanie i długi czas prowadzenia reakcji, a produkt może powstawać w różnych fragmentach profilu temperaturowego, zależnie od reaktywności substratu. Przypisanie każdej reakcji jednej efektywnej temperatury wprowadzałoby więc założenie, którego protokół doświadczalny nie uzasadnia.

Jest to różnica między danymi tworzonymi pod modelowanie a danymi powstającymi w zwykłej praktyce laboratoryjnej. W kampanii zaprojektowanej pod uczenie maszynowe temperaturę utrzymuje się stałą właśnie po to, by przeliczenie było poprawne. W laboratorium chemii organicznej reakcję optymalizuje się natomiast pod kątem enancjoselektywności, a nie pod kątem wiarygodności późniejszego przeliczenia energetycznego. Pozostaliśmy zatem przy %*ee*.

6.2.6 Walidacja eksperymentalna

Praktyczną użyteczność modeli sprawdziłam na substratach nieobecnych w zbiorach treningowych, we współpracy z dr J. Jaszczewską-Adamczak, która wykonała część eksperymentów epoksydacji oraz analizy dla otrzymanych epoksydów. Do walidacji wytypowałam 36 dodatkowych struktur, które najpierw oceniono klasyfikatorem, a następnie przebadałam doświadczalnie w standardowych warunkach (Rysunek 32).

Klasyfikator poprawnie rozpoznał znaczną część zarówno substratów reaktywnych, jak i niereaktywnych, wystąpiły jednak przypadki fałszywie pozytywne (Rysunek 33). Substraty niosące silnie elektronoakceptorowe lub podstawniki powodujące zwiększenie zawady przestrzennej były konsekwentnie i poprawnie klasyfikowane jako niereaktywne, w pełnej zgodzie z obserwacją. Wyniki fałszywie pozytywne są przy tym cenną informacją, ponieważ wyznaczają granicę klasyfikatora i wskazują obszary, w których podobieństwo strukturalne do substratów produktywnych nie wystarcza, aby zapewnić reaktywność. Klasyfikacja stanowi tutaj użyteczny filtr pierwszego etapu.

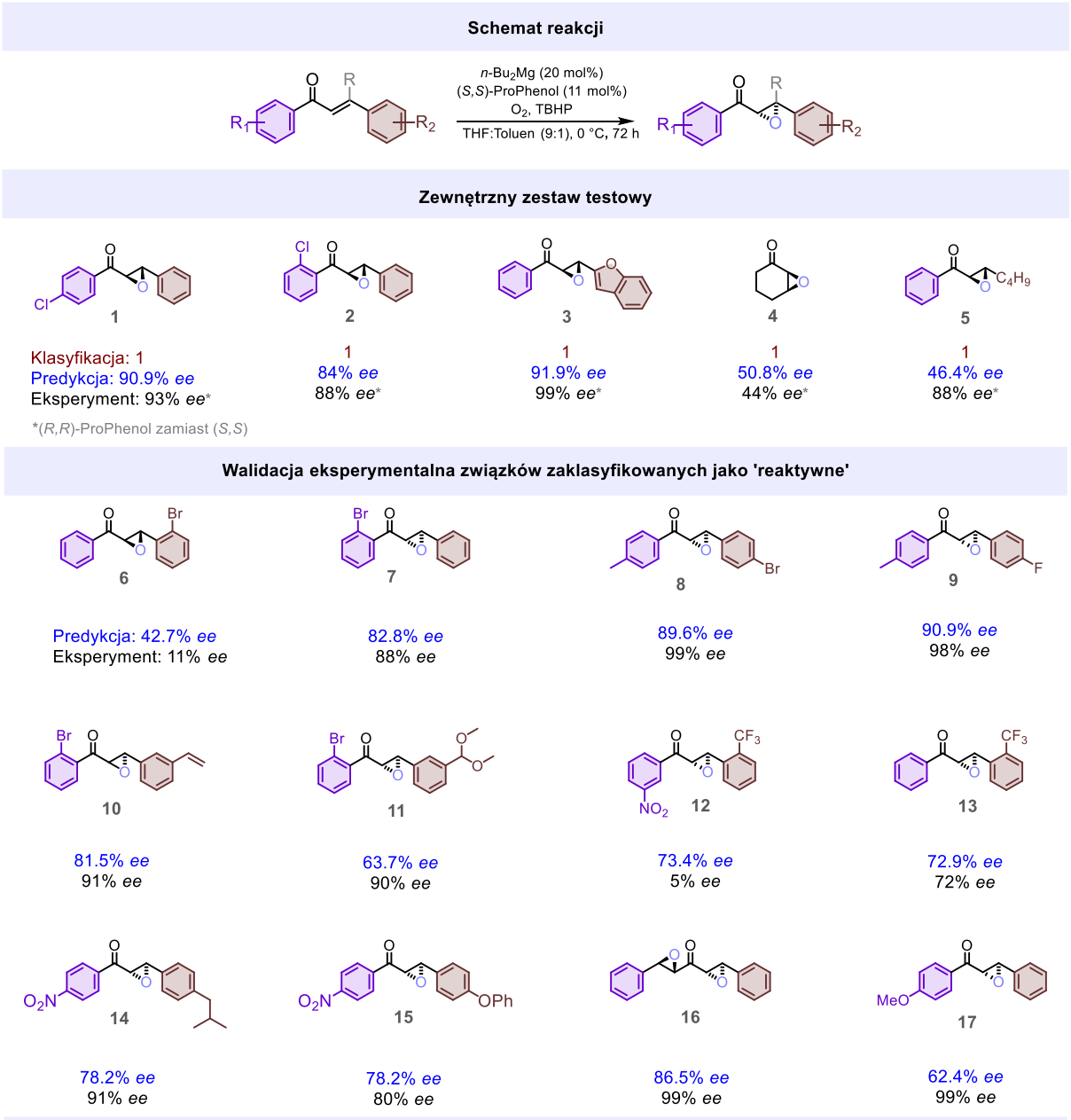


***Rysunek 32**. Klasyfikacja reaktywności dla 36 substratów walidowanych prospektywnie, z podziałem na substraty reaktywne i niereaktywne.*

6.2.7 Zakres stosowalności i granice predykcji

Analizę zakresu stosowalności, przygotowaną przez mgr. D. Nowaka, wykorzystałam do rozróżnienia predykcji interpolacyjnych i ekstrapolacyjnych (dostępne w materiałach uzupełniających pracy). Podział ten nie służył do formalnego odrzucania predykcji, lecz do oceny ich wiarygodności i wskazania wyników, które wymagają potwierdzenia eksperymentalnego przed podjęciem decyzji syntetycznej.

Najbardziej pouczający był przypadek związku **12** (Rysunek 33). Substrat zawiera dwa silnie elektronoakceptorowe podstawniki, grupę nitrową w pozycji 3 pierścienia arylowego sąsiadującego z grupą karbonylową oraz grupę CF_3 w pozycji orto drugiego pierścienia. Model przewidział dla niego 73,4% *ee*, podczas gdy eksperymentalnie otrzymano 5% *ee*. Jednocześnie analiza zakresu stosowalności umieściła ten związek poza obszarem dobrze reprezentowanym przez dane treningowe. W zbiorze treningowym nie występowały substraty łączące tak silne efekty elektronoakceptorowe po obu stronach układu enonowego. Przypadek ten pokazuje, że wartość predykcji należy interpretować łącznie z informacją o położeniu substratu względem zakresu stosowalności, ponieważ sama liczba nie informuje o tym, czy model przewiduje w obrębie znanej przestrzeni chemicznej, czy poza nią.



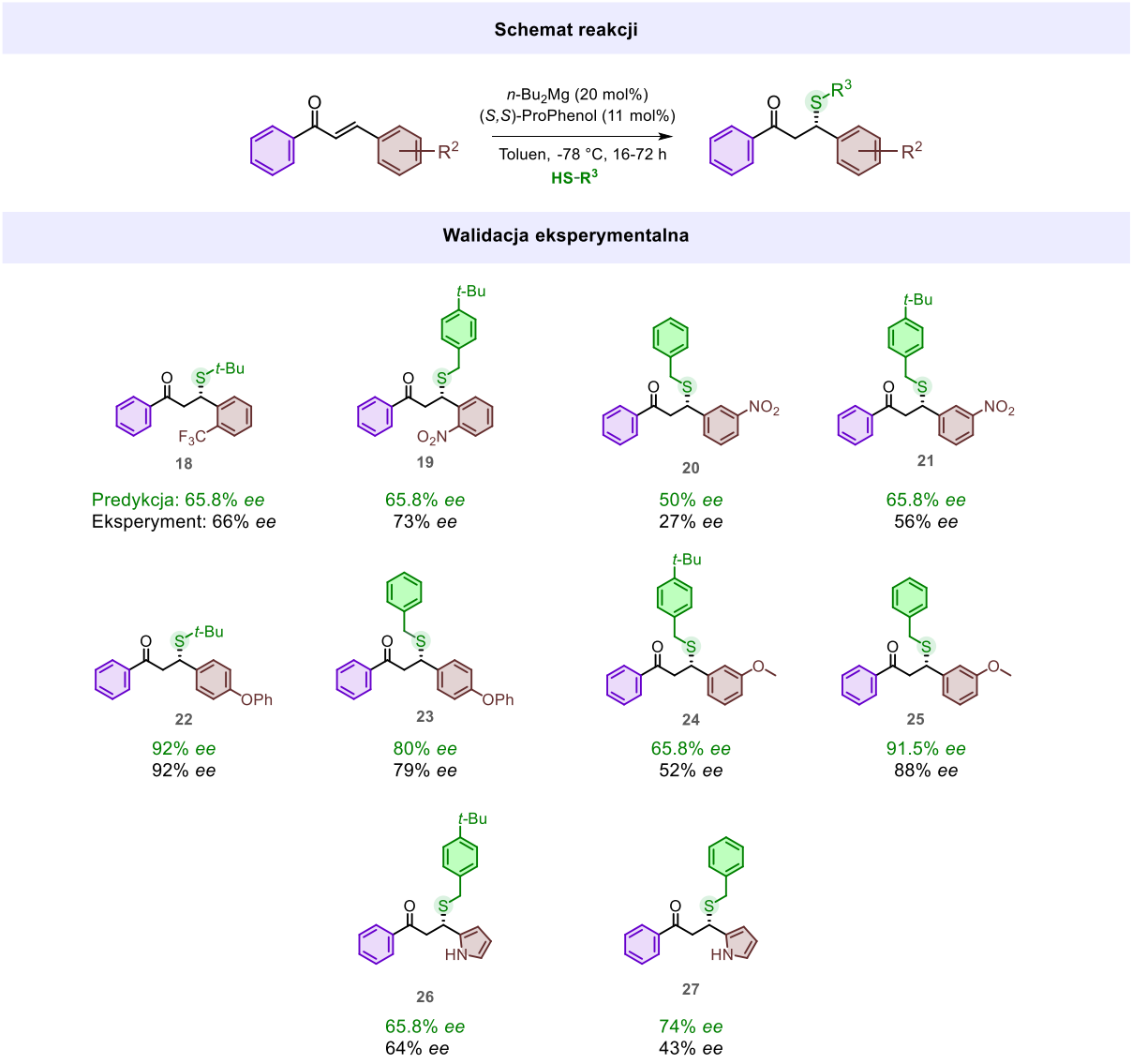
Rysunek 33. *Walidacja eksperymentalna substratów przewidzianych jako reaktywne wraz z porównaniem wartości przewidzianych i zmierzonych.*

6.2.8 Addycja tio-Michaela

Zbiór addycji tio-Michaela, obejmujący dwa niezależnie zmieniane substraty, pozwolił zbudować model regresyjny o mierzalnej wartości predykcyjnej. Modelem końcowym, po selekcji cech prowadzonej w schemacie k krotnym, okazało się drzewo decyzyjne oparte wyłącznie na deskryptorach Mordred[31]. Model zbudowano na 63 reakcjach, z których 53 wykorzystano do treningu, a 10 do testu. Prospektywny zbiór walidacyjny obejmował wyłącznie kombinacje substratów niewykorzystywane na żadnym etapie budowy modelu.

Pierwsze modele miały ograniczoną zdolność do generalizacji. Potraktowałam ten wynik jako informację o niedostatecznym pokryciu przestrzeni chemicznej, a nie jako podstawę do dalszego dostrajania modelu do konkretnego podziału danych. Zdecydowaliśmy więc o wykonaniu kolejnych reakcji w tych samych warunkach, aby zwiększyć zróżnicowanie zbioru. Substraty do tych doświadczeń przygotowała dr Joanna Jaszczewska Adamczak, natomiast ja przeprowadziłam reakcje walidacyjne (Rysunek 34), analizę produktów oraz interpretację uzyskanych wyników. Część wcześniej wyłączonych reakcji została następnie włączona do zmodyfikowanego zbioru, dlatego nie traktowano ich jako niezależnej walidacji modelu końcowego, lecz jako element dalszego opracowywania danych.

Powtarzające się wartości przewidywane dla części substratów walidacyjnych wynikają z budowy drzewa decyzyjnego. Chemicznie podobne substraty mogą trafiać do tego samego ‘liścia’ i otrzymywać tę samą wartość predykcji. Z tego względu model oceniano nie tylko na podstawie błędu bezwzględnego, ale również pod kątem zdolności do porządkowania kombinacji substratów według oczekiwanej enancjoselektywności. W tej roli model okazał się użyteczny jako narzędzie pomagające ustalić, które kombinacje warto sprawdzić w pierwszej kolejności.



Rysunek 34. *Reakcja tio-Michaela: schemat reakcji oraz walidacja eksperymentalna predykcji enancjoselektywności wraz z porównaniem przewidywanych i obserwowanych wartości nadmiaru enancjomerycznego. Przewidywane wartości ee przedstawiono z zaokrągleniem zgodnym z dokładnością zastosowaną na rysunku. Wartości MAE, RMSE i R^2 obliczono na podstawie niezaokrąglonych wyników modelu.*

6.2.9 Podsumowanie

Dla stosunkowo jednorodnego zbioru epoksydacji chalkonów najbardziej wiarygodnym i praktycznie użytecznym wynikiem okazała się klasyfikacja reaktywności. Regresja enancjoselektywności była natomiast ograniczona przez skupienie większości wyników w zakresie wysokich wartości *ee* oraz dużą wrażliwość na niewielkie zmiany składu zbioru. Bardziej zróżnicowany zbiór addycji tio-Michaela pozwolił zbudować model regresyjny o mierzalnej wartości predykcyjnej. Jego główna użyteczność polegała jednak na porządkowaniu kombinacji substratów według oczekiwanej selektywności, a nie na dokładnym przewidywaniu bezwzględnych wartości *ee*.

Walidacja prospektywna pokazała, że modele wychwytują zależności pomiędzy strukturą a reaktywnością i enancjoselektywnością, ale jednocześnie ujawniła dużą wrażliwość modeli budowanych na małych zbiorach na ich skład. Szczególnie istotne okazały się przypadki nieudanych predykcji. Pokazały one, że wartość przewidywana powinna być interpretowana razem z zakresem stosowalności modelu, ponieważ dla substratów leżących poza przestrzenią reprezentowaną przez dane treningowe model może zwracać pozornie wiarygodną liczbę mimo braku odpowiednich przykładów do jej wyznaczenia.

Wyniki tej pracy pokazują, że w małych zbiorach z katalizy asymetrycznej o użyteczności modelu decyduje przede wszystkim sposób przygotowania danych, właściwe sformułowanie pytania oraz rzetelna walidacja. Klasyczne zakresy stosowalności metod katalitycznych powstają w celu stwierdzenia jak uniwersalna lub ograniczona jest metoda, a nie stworzenia zrównoważonego zbioru do modelowania. To właśnie struktura dostępnych danych określa więc, jakie pytania można modelowi wiarygodnie postawić. W takim ujęciu uczenie maszynowe nie zastępuje eksperymentu, lecz pomaga wskazać, które substraty warto badać w pierwszej kolejności oraz kiedy przewidywanie wymaga szczególnej ostrożności.

Badania opisane w tym rozdziale zostały opublikowane w pracy: P. Baczewska, D. Nowak, J. A. Jaszczewska-Adamczak, R. A. Bachorz, M. Hoffmann, J. Mlynarski, Predicting Reactivity and Enantioselectivity from Small Experimental Datasets: A Case Study in Asymmetric Magnesium Catalysis, ACS Catal., 2026; https://doi.org/10.1021/acscatal.6c05090

Szczegółowy podział wkładu przedstawiają oświadczenia współautorów.

7. Podsumowanie i wnioski

Obie prace pokazują, że rodzaj dostępnych danych wyznacza zakres pytań, które można wiarygodnie postawić modelowi. Dane literaturowe, po ręcznym opracowaniu i ujednoliceniu, pozwoliły zbudować model rekomendujący kompleksy magnezu wykorzystywane w katalizie asymetrycznej. Jego praktyczną wartość sprawdziłam eksperymentalnie. W trzech redukcjach α-ketoestrów rekomendowany układ $Mg(OTf)_2$/**L-2** dawał wyższe *ee* niż układ literaturowy. Dla γ,β-nienasyconego ketoestru oraz izatyny wzbogaciły ilość dostępnych asymetrycznych redukcji katalizowanych magnezem, których wcześniej dla tych klas substratów nie opisano. W przypadku związku prowadzącego do prekursora BMS-270394 uzyskałam do 89% *ee* przy wysokiej wydajności, co pokazało, że rekomendacje modelu mogą stanowić rzeczywisty punkt wyjścia do dalszej optymalizacji syntetycznej.

Małe, jednorodne zbiory laboratoryjne pozwoliły odpowiedzieć na inne pytania. Dla epoksydacji chalkonów najbardziej użyteczna okazała się klasyfikacja reaktywności, natomiast dla bardziej zróżnicowanej addycji tio-Michaela możliwe było uporządkowanie kombinacji substratów według oczekiwanej enancjoselektywności. Predykcje sprawdziłam prospektywnie na nowych substratach. Analiza zakresu stosowalności pozwoliła również wskazać przypadki, w których model wychodził poza dobrze reprezentowaną przestrzeń chemiczną i jego wynik wymagał szczególnej ostrożności.

Za najważniejsze osiągnięcia projektu uważam:

- zbudowanie największej i najdokładniejszej bazy asymetrycznych reakcji katalizowanych kompleksami magnezu, obejmującej 1538 reakcji, 139 ligandów chiralnych i 104 typy reakcji. Istotnym elementem było odzyskanie z materiałów uzupełniających również wyników słabych, w tym 128 reakcji o *ee* poniżej 50%, które są słabo reprezentowane w typowych bazach reakcji.
- pokazanie, że rozproszone i niejednorodne dane literaturowe mogą zostać wykorzystane do rekomendowania układów katalitycznych i rzeczywiście wspierać pierwszy etap optymalizacji reakcji. Najważniejszym potwierdzeniem były wyniki eksperymentalne, w tym poprawa enancjoselektywności znanych redukcji α-ketoestrów oraz uzyskanie nowych przykładów asymetrycznych redukcji katalizowanych magnezem.
- określenie ograniczeń modeli wynikających nie tylko z liczby danych, ale przede wszystkim z ich struktury. Na przykładzie epoksydacji pokazałam, że brak produktu nie może być traktowany jako 0% *ee* i że rozdzielenie reaktywności od enancjoselektywności może być konieczne, aby postawić chemicznie poprawne zadanie predykcyjne.
- zbadanie, w jakim zakresie małe, ale dobrze kontrolowane zbiory laboratoryjne mogą dostarczać użytecznych lokalnych informacji o reaktywności i enancjoselektywności. Wyniki pokazały przy tym znaczenie prospektywnej walidacji oraz kontroli zakresu stosowalności modelu.

Pokazałam, że zarówno rozproszone dane literaturowe, jak i niewielkie zbiory laboratoryjne mogą dostarczać chemikowi praktycznie użytecznej informacji, pod warunkiem że modelowi stawia się pytanie odpowiadające temu, czego może się z tych danych nauczyć. W takim ujęciu model nie zastępuje eksperymentu. Pomaga natomiast wybrać, co warto sprawdzić w pierwszej kolejności, i rozpoznać sytuacje, w których jego przewidywaniu nie należy ufać.

Ograniczenia takich modeli są nadal wyraźne i wyniki tej rozprawy pokazują również, że nie każdą przestrzeń reakcji można równie dobrze opisać na podstawie dostępnych danych. Jednocześnie metody uczenia maszynowego i sztucznej inteligencji rozwijają się wyjątkowo

szybko. Strategie uznawane dziś za obiecujące mogą w krótkim czasie zostać zastąpione przez rozwiązania skuteczniejsze, dlatego trudno obecnie wskazać jeden kierunek, który stanie się standardem projektowania katalizatorów.

Liczba dostępnych danych chemicznych rośnie szybciej niż możliwość ich manualnego przeanalizowania i połączenia w jeden spójny obraz. W tym właśnie widzę szczególnie uzasadnione miejsce dla odpowiedzialnego wykorzystania uczenia maszynowego i sztucznej inteligencji w nauce. Nie jako zastępstwa dla wiedzy chemicznej, interpretacji ani eksperymentu, lecz jako narzędzia pozwalającego porządkować coraz większą ilość informacji i dostrzegać zależności, których ręczne odnalezienie staje się coraz trudniejsze. W tej swoistej „klęsce urodzaju" danych największą wartością takich metod może być nie automatyzacja samego odkrycia, ale pomoc chemikowi w zobaczeniu szerszego obrazu i wyborze eksperymentu, który rzeczywiście warto wykonać.

Bibliografia

(1) Yang, H.; Yu, H.; Stolarzewicz, I. A.; Tang, W. Enantioselective Transformations in the Synthesis of Therapeutic Agents. *Chem. Rev.* **2023**, *123* (15), 9397–9446. https://doi.org/10.1021/acs.chemrev.3c00010.

(2) Sharma, S. K.; Paniraj, A. S. R.; Tambe, Y. B. Developments in the Catalytic Asymmetric Synthesis of Agrochemicals and Their Synthetic Importance. *J. Agric. Food Chem.* **2021**, *69* (49), 14761–14780. https://doi.org/10.1021/acs.jafc.1c05553.

(3) Meskers, S. C. J. Consequences of chirality on the response of materials. *Mater. Adv.* **2022**, *3* (5), 2324–2336. https://doi.org/10.1039/D1MA01053K.

(4) Senkuttuvan, N.; Komarasamy, B.; Krishnamoorthy, R.; Sarkar, S.; Dhanasekaran, S.; Anaikutti, P. The significance of chirality in contemporary drug discovery-a mini review. *RSC Adv.* **2024**, *14* (45), 33429–33448. https://doi.org/10.1039/D4RA05694A.

(5) McVicker, R. U.; O'Boyle, N. M. Chirality of New Drug Approvals (2013–2022): Trends and Perspectives. *J. Med. Chem.* **2024**, *67* (4), 2305–2320. https://doi.org/10.1021/acs.jmedchem.3c02239.

(6) Nie, L.-D.; Shi, X.-X.; Ko, K. H.; Lu, W.-D. A Short and Practical Synthesis of Oseltamivir Phosphate (Tamiflu) from (–)-Shikimic Acid. *J. Org. Chem.* **2009**, *74* (10), 3970–3973. https://doi.org/10.1021/jo900218k.

(7) Boschi, F.; Camps, P.; Comes-Franchini, M.; Muñoz-Torrero, D.; Ricci, A.; Sánchez, L. A synthesis of levetiracetam based on (S)-N-phenylpantolactam as a chiral auxiliary. *Tetrahedron: Asymmetry* **2005**, *16* (22), 3739–3745. https://doi.org/10.1016/j.tetasy.2005.10.014.

(8) Martinez, C. A.; Hu, S.; Dumond, Y.; Tao, J.; Kelleher, P.; Tully, L. Development of a Chemoenzymatic Manufacturing Process for Pregabalin. *Org. Process Res. Dev.* **2008**, *12* (3), 392–398. https://doi.org/10.1021/op7002248.

(9) Hansen, K. B.; Hsiao, Y.; Xu, F.; Rivera, N.; Clausen, A.; Kubryk, M.; Krska, S.; Rosner, T.; Simmons, B.; Balsells, J.; Ikemoto, N.; Sun, Y.; Spindler, F.; Malan, C.; Grabowski, E. J. J.; Armstrong, J. D. Highly Efficient Asymmetric Synthesis of Sitagliptin. *J. Am. Chem. Soc.* **2009**, *131* (25), 8798–8804. https://doi.org/10.1021/ja902462q.

(10) Sheldon, R. A. Chirotechnology: Designing Economic Chiral Syntheses. *J. Chem. Technol. Biotechnol.* **1996**, *67* (1), 1–14. https://doi.org/10.1002/(SICI)1097-4660(199609)67:1<1::AID-JCTB531>3.0.CO;2-L.

(11) The Nobel Prize in Chemistry 2001. NobelPrize.org. https://www.nobelprize.org/prizes/chemistry/2001/summary/.

(12) The Nobel Prize in Chemistry 2021. NobelPrize.org. https://www.nobelprize.org/prizes/chemistry/2021/summary/.

(13) Ruth, M.; Gensch, T.; Schreiner, P. R. Contrasting Historical and Physical Perspectives in Asymmetric Catalysis: ΔΔG≠ versus Enantiomeric Excess. *Angew. Chem., Int. Ed.* **2024**, *63* (48), e202410308. https://doi.org/10.1002/anie.202410308.

(14) Tsuji, N.; Sidorov, P.; Zhu, C.; Nagata, Y.; Gimadiev, T.; Varnek, A.; List, B. Predicting Highly Enantioselective Catalysts Using Tunable Fragment Descriptors. *Angew. Chem., Int. Ed.* **2023**, *62* (11), e202218659. https://doi.org/10.1002/anie.202218659.

(15) Hammett, L. P. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, *59* (1), 96–103. https://doi.org/10.1021/ja01280a022.

(16) 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.* **2016**, *49* (6), 1292–1301. https://doi.org/10.1021/acs.accounts.6b00194.

(17) Shibasaki, M.; Sasai, H.; Arai, T. Asymmetric Catalysis with Heterobimetallic Compounds. *Angew. Chem., Int. Ed. Engl.* **1997**, *36* (12), 1236–1256. https://doi.org/10.1002/anie.199712361.

(18) Pietrzak, T.; Kubisiak, M.; Justyniak, I.; Zelga, K.; Bojarski, E.; Tratkiewicz, E.; Ochal, Z.; Lewiński, J. Oxygenation Chemistry of Magnesium Alkyls Incorporating β-Diketiminate Ligands Revisited. *Chem. - Eur. J.* **2016**, *22* (49), 17776–17783. https://doi.org/10.1002/chem.201603931.

(19) Yang, D.; Wang, L.; Li, D.; Wang, R. Magnesium Catalysis in Asymmetric Synthesis. *Chem* **2019**, *5* (5), 1108–1166. https://doi.org/10.1016/j.chempr.2019.02.002.

(20) Czombik, A. M.; Gajewy, J.; Czapik, A.; Kwit, M. Magnesium-catalyzed stereoselective transformations – A survey through recent achievements. *Polyhedron* **2022**, *219*, 115790. https://doi.org/10.1016/j.poly.2022.115790.

(21) Mudd, G. M.; Jowitt, S. M.; Werner, T. T. Global platinum group element resources, reserves and mining – A critical assessment. *Sci. Total Environ.* **2018**, *622–623*, 614–625. https://doi.org/10.1016/j.scitotenv.2017.11.350.

(22) Hughes, A. E.; Haque, N.; Northey, S. A.; Giddey, S. Platinum Group Metals: A Review of Resources, Production and Usage with a Focus on Catalysts. *Resources* **2021**, *10* (9), 93. https://doi.org/10.3390/resources10090093.

(23) Bao, D. Dynamics and correlation of platinum-group metals spot prices. *Resour. Policy* **2020**, *68*, 101772. https://doi.org/10.1016/j.resourpol.2020.101772.

(24) International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). *ICH Harmonised Guideline: Guideline for Elemental Impurities Q3D(R2)*; Step 4, 26 April 2022. https://database.ich.org/sites/default/files/Q3D-R2_Guideline_Step4_2022_0308.pdf.

(25) Liang, Y.; Luo, J.; Diskin-Posner, Y.; Milstein, D. Designing New Magnesium Pincer Complexes for Catalytic Hydrogenation of Imines and N-Heteroarenes: H_2 and N–H Activation by Metal–Ligand Cooperation as Key Steps. *J. Am. Chem. Soc.* **2023**, *145* (16), 9164–9175. https://doi.org/10.1021/jacs.3c01091.

(26) Dalmau, D.; García-Abellán, S.; Alegre-Requena, J. V. Machine Learning in Homogeneous Catalysis: Basic Concepts and Best Practices. *ACS Catal.* **2026**, *16* (1), 1–11. https://doi.org/10.1021/acscatal.5c06439.

(27) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754. https://doi.org/10.1021/ci100050t.

(28) Landrum, G.; Tosco, P.; Kelley, B.; Rodriguez, R.; Cosgrove, D.; Vianello, R.; et al. *RDKit: Open-Source Cheminformatics Software*, Release 2025_09_3 (Q3 2025); Zenodo, 2025. https://doi.org/10.5281/zenodo.17746401.

(29) Gedeck, P.; Rohde, B.; Bartels, C. QSAR – How Good Is It in Practice? Comparison of Descriptor Sets on an Unbiased Cross Section of Corporate Data Sets. *J. Chem. Inf. Model.* **2006**, *46* (5), 1924–1936. https://doi.org/10.1021/ci050413p.

(30) Byadi, S.; Gantzer, P.; Gimadiev, T.; Sidorov, P. DOPtools: a Python platform for descriptor calculation and model optimization. *Digital Discovery* **2025**, *4* (5), 1188–1198. https://doi.org/10.1039/D4DD00399C.

(31) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminf.* **2018**, *10* (1), 4. https://doi.org/10.1186/s13321-018-0258-y.

(32) Gensch, T.; dos Passos Gomes, G.; Friederich, P.; Peters, E.; Gaudin, T.; Pollice, R.; Jorner, K.; Nigam, A.; Lindner-D'Addario, M.; Sigman, M. S.; Aspuru-Guzik, A. A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis. *J. Am. Chem. Soc.* **2022**, *144* (3), 1205–1217. https://doi.org/10.1021/jacs.1c09718.

(33) Harper, K. C.; Bess, E. N.; Sigman, M. S. Multidimensional steric parameters in the analysis of asymmetric catalytic reactions. *Nat. Chem.* **2012**, *4* (5), 366–374. https://doi.org/10.1038/nchem.1297.

(34) Brethomé, A. V.; Fletcher, S. P.; Paton, R. S. Conformational Effects on Physical-Organic Descriptors: The Case of Sterimol Steric Parameters. *ACS Catal.* **2019**, *9* (3), 2313–2323. https://doi.org/10.1021/acscatal.8b04043.

(35) Tolman, C. A. Steric effects of phosphorus ligands in organometallic chemistry and homogeneous catalysis. *Chem. Rev.* **1977**, *77* (3), 313–348. https://doi.org/10.1021/cr60307a002.

(36) Clavier, H.; Nolan, S. P. Percent buried volume for phosphine and N-heterocyclic carbene ligands: steric properties in organometallic chemistry. *Chem. Commun.* **2010**, *46* (6), 841. https://doi.org/10.1039/b922984a.

(37) Cammarota, R. C.; Liu, W.; Bacsa, J.; Davies, H. M. L.; Sigman, M. S. Mechanistically Guided Workflow for Relating Complex Reactive Site Topologies to Catalyst Performance in C–H Functionalization Reactions. *J. Am. Chem. Soc.* **2022**, *144* (4), 1881–1898. https://doi.org/10.1021/jacs.1c12198.

(38) Baczewska, P.; Kulczykowski, M.; Zambroń, B.; Jaszczewska-Adamczak, J.; Pakulski, Z.; Roszak, R.; Grzybowski, B. A.; Mlynarski, J. Machine Learning Algorithm Guides Catalyst Choices for Magnesium-Catalyzed Asymmetric Reactions. *Angew. Chem., Int. Ed.* **2024**, *63* (37), e202318487. https://doi.org/10.1002/anie.202318487.

(39) Zahrt, A. F.; Athavale, S. V.; Denmark, S. E. Quantitative Structure–Selectivity Relationships in Enantioselective Catalysis: Past, Present, and Future. *Chem. Rev.* **2020**, *120* (3), 1620–1689. https://doi.org/10.1021/acs.chemrev.9b00425.

(40) Chuang, K. V.; Keiser, M. J. Comment on "Predicting reaction performance in C–N cross-coupling using machine learning". *Science* **2018**, *362* (6416), eaat8603. https://doi.org/10.1126/science.aat8603.

(41) LeSueur, A.; Bianchi, P.; Gallarati, S.; Lefave, S. J.; Sigman, M. S. Best Practices and Considerations for Applying Multiple Linear Regression in Organic Chemistry Research. *J. Org. Chem.* **2026**, *91* (17), 5733–5744. https://doi.org/10.1021/acs.joc.5c03206.

(42) Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*; 2017. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

(43) Moldagulov, G.; Lee, K.; Nurgaliyev, S.; Salem, A.; Kuznietsov, A.; Grzybowski, B. A. Hybrid Computational Strategy for Predicting Complex Ligand–Metal Architectures. *Angew. Chem., Int. Ed.* **2026**, *65* (14), e24655. https://doi.org/10.1002/anie.202524655.

(44) 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.* **1988**, *110* (18), 5959–5967. https://doi.org/10.1021/ja00226a005.

(45) Verma, J.; Khedkar, V.; Coutinho, E. 3D-QSAR in Drug Design - A Review. *Curr. Top. Med. Chem.* **2010**, *10* (1), 95–115. https://doi.org/10.2174/156802610790232260.

(46) 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.* **2003**, *68* (12), 4648–4656. https://doi.org/10.1021/jo0267697.

(47) 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.* **2019**, *92* (10), 1701–1706. https://doi.org/10.1246/bcsj.20190132.

(48) Milo, A.; Neel, A. J.; Toste, F. D.; Sigman, M. S. A data-intensive approach to mechanistic elucidation applied to chiral anion catalysis. *Science* **2015**, *347* (6223), 737–743. https://doi.org/10.1126/science.1261043.

(49) 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* **2018**, *362* (6415), 670–674. https://doi.org/10.1126/science.aat2299.

(50) Reid, J. P.; Sigman, M. S. Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature* **2019**, *571* (7765), 343–348. https://doi.org/10.1038/s41586-019-1384-z.

(51) 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* **2019**, *363* (6424), eaau5631. https://doi.org/10.1126/science.aau5631.

(52) Betinol, I. O.; Lai, J.; Thakur, S.; Reid, J. P. A Data-Driven Workflow for Assigning and Predicting Generality in Asymmetric Catalysis. *J. Am. Chem. Soc.* **2023**, *145* (23), 12870–12883. https://doi.org/10.1021/jacs.3c03989.

(53) Gallarati, S.; Fabregat, R.; Laplaza, R.; Bhattacharjee, S.; Wodrich, M. D.; Corminboeuf, C. Reaction-based machine learning representations for predicting the enantioselectivity of organocatalysts. *Chem. Sci.* **2021**, *12* (20), 6879–6889. https://doi.org/10.1039/d1sc00482d.

(54) Schoepfer, A. A.; Laplaza, R.; Wodrich, M. D.; Waser, J.; Corminboeuf, C. Reaction-Agnostic Featurization of Bidentate Ligands for Bayesian Ridge Regression of Enantioselectivity. *ACS Catal.* **2024**, *14* (12), 9302–9312. https://doi.org/10.1021/acscatal.4c02452.

(55) Gallarati, S.; Bucci, E. M.; Doyle, A. G.; Sigman, M. S. Transferable enantioselectivity models from sparse data. *Nature* **2026**, *651* (8106), 637–646. https://doi.org/10.1038/s41586-026-10239-7.

(56) 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* **2018**, *360* (6385), 186–190. https://doi.org/10.1126/science.aar5169.

(57) Rinehart, N. I.; Saunthwal, R. K.; Wellauer, J.; Zahrt, A. F.; Schlemper, L.; Shved, A. S.; Bigler, R.; Fantasia, S.; Denmark, S. E. A machine-learning tool to predict substrate-adaptive conditions

for Pd-catalyzed C–N couplings. *Science* **2023**, *381* (6661), 965–972. https://doi.org/10.1126/science.adg2114.

(58) Olen, C. L.; Zahrt, A. F.; Reilly, S. W.; Schultz, D.; Emerson, K.; Candito, D.; Wang, X.; Strotman, N. A.; Denmark, S. E. Chemoinformatic Catalyst Selection Methods for the Optimization of Copper–Bis(oxazoline)-Mediated, Asymmetric, Vinylogous Mukaiyama Aldol Reactions. *ACS Catal.* **2024**, *14* (4), 2642–2655. https://doi.org/10.1021/acscatal.3c05903.

(59) Dotson, J. J.; van Dijk, L.; Timmerman, J. C.; Grosslight, S.; Walroth, R. C.; Gosselin, F.; Püntener, K.; Mack, K. A.; Sigman, M. S. Data-Driven Multi-Objective Optimization Tactics for Catalytic Asymmetric Reactions Using Bisphosphine Ligands. *J. Am. Chem. Soc.* **2023**, *145* (1), 110–121. https://doi.org/10.1021/jacs.2c08513.

(60) Xu, L.-C.; Zhang, S.-Q.; Li, X.; Tang, M.-J.; Xie, P.-P.; Hong, X. Towards Data-Driven Design of Asymmetric Hydrogenation of Olefins: Database and Hierarchical Learning. *Angew. Chem., Int. Ed.* **2021**, *60* (42), 22804–22811. https://doi.org/10.1002/anie.202106880.

(61) Tang, M.-J.; Zhang, T.; Huang, Q.; Li, S.; Liu, R.; Li, H.; Chen, X.; Dong, S.; Liu, X.; Feng, X.; Hong, X. Data-Driven Modeling of N,N′-Dioxide/Metal-Catalyzed Asymmetric Michael Additions. *Angew. Chem., Int. Ed.* **2026**, *65* (12), e18560. https://doi.org/10.1002/anie.202518560.

(62) Singh, S.; Hernández-Lobato, J. M. A meta-learning approach for selectivity prediction in asymmetric catalysis. *Nat. Commun.* **2025**, *16* (1), 3599. https://doi.org/10.1038/s41467-025-58854-8.

(63) Ocampo, B. E.; Altundas, B.; Bock, M. J.; Feiz, S.; Denmark, S. E. Data-Driven Prediction of Enantioselectivity for the Sharpless Asymmetric Dihydroxylation: Model Development and Experimental Validation. *ACS Cent. Sci.* **2025**, *11* (9), 1640–1650. https://doi.org/10.1021/acscentsci.5c00900.

(64) Beker, W.; Roszak, R.; Wołos, A.; Angello, N. H.; Rathore, V.; Burke, M. D.; Grzybowski, B. A. Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki–Miyaura Coupling. *J. Am. Chem. Soc.* **2022**, *144* (11), 4819–4827. https://doi.org/10.1021/jacs.1c12005.

(65) Hoque, A.; Jain, N.; Chenna, D.; Sunoj, R. B. Machine learning models for catalytic asymmetric reactions of simple alkenes: from enantioselectivity predictions to chemical insights. *Digital Discovery* **2026**, *5* (4), 1718–1735. https://doi.org/10.1039/d5dd00483g.

(66) Lee, K.; Moldagulov, G.; Grzybowski, B. A. The Fragility of Bioactivity Prediction: Rigorous Dataset Splits Expose the Illusion of ML Accuracy. *Chem. - Eur. J.* **2026**, e71208. https://doi.org/10.1002/chem.71208.

(67) Gao, W.; Raghavan, P.; Shprints, R.; Coley, C. W. Revealing the Relationship between Publication Bias and Chemical Reactivity with Contrastive Learning. *J. Am. Chem. Soc.* **2025**, *147* (10), 8959–8968. https://doi.org/10.1021/jacs.5c01120.

(68) Vogl, E. M.; Gröger, H.; Shibasaki, M. Towards Perfect Asymmetric Catalysis: Additives and Cocatalysts. *Angew. Chem., Int. Ed.* **1999**, *38* (11), 1570–1577. https://doi.org/10.1002/(SICI)1521-3773(19990601)38:11<1570::AID-ANIE1570>3.0.CO;2-Y.

(69) Hong, L.; Sun, W.; Yang, D.; Li, G.; Wang, R. Additive Effects on Asymmetric Catalysis. *Chem. Rev.* **2016**, *116* (6), 4006–4123. https://doi.org/10.1021/acs.chemrev.5b00676.

(70) Swain, M. MolVS: Molecule Validation and Standardization. https://molvs.readthedocs.io/.

(71) Szymkuć, S.; Badowski, T.; Grzybowski, B. A. Is Organic Chemistry Really Growing Exponentially? *Angew. Chem., Int. Ed.* **2021**, *60* (50), 26226–26232. https://doi.org/10.1002/anie.202111540.

(72) Stenutz, R. Tables for Organic Chemistry. http://www.stenutz.eu/chem/.

(73) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; ACM: New York, NY, USA, 2019; pp 2623–2631. https://doi.org/10.1145/3292500.3330701.

(74) Probst, D.; Schwaller, P.; Reymond, J.-L. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digital Discovery* **2022**, *1* (2), 91–97. https://doi.org/10.1039/d1dd00006c.

(75) Enders, D.; Stöckel, B. A.; Rembiak, A. Enantio- and chemoselective Brønsted-acid/$Mg(nBu)_2$ catalysed reduction of α-keto esters with catecholborane. *Chem. Commun.* **2014**, *50* (34), 4489–4491. https://doi.org/10.1039/c4cc00427b.

(76) Li, J.; Liao, Y.; Zhang, Y.; Liu, X.; Lin, L.; Feng, X. Chiral magnesium(II)-catalyzed asymmetric ring-opening of meso-aziridines with primary alcohols. *Chem. Commun.* **2014**, *50* (50), 6672–6674. https://doi.org/10.1039/c4cc02206h.

(77) Yin, C.; Cao, W.; Lin, L.; Liu, X.; Feng, X. N,N′-Dioxide-Magnesium Ditriflate Complex-Catalyzed Asymmetric α-Hydroxylation of β-Keto Esters and β-Keto Amides. *Adv. Synth. Catal.* **2013**, *355* (10), 1924–1930. https://doi.org/10.1002/adsc.201300335.

(78) Xiao, W.; Ning, L.; Xin, S.; Dong, S.; Liu, X.; Feng, X. Enantioselective [2+2] Cycloaddition of Allenyl Imide with Mono- or Disubstituted Alkenes. *Angew. Chem., Int. Ed.* **2022**, *61* (44), e202211596. https://doi.org/10.1002/anie.202211596.

(79) Qiu, H.; Ren, J.; Zhang, L.; Song, R.; Si, W.; Yang, D.; Wen, L.; Lv, J. Enantioselective Meerwein–Ponndorf–Verley reduction of β,γ-unsaturated α-keto esters by asymmetric binary-acid catalysis in the green solvent iPrOH. *Green Chem.* **2023**, *25* (10), 3948–3955. https://doi.org/10.1039/d2gc03441g.

(80) Vasilenko, V.; Blasius, C. K.; Wadepohl, H.; Gade, L. H. Borohydride intermediates pave the way for magnesium-catalysed enantioselective ketone reduction. *Chem. Commun.* **2020**, *56* (8), 1203–1206. https://doi.org/10.1039/c9cc09111d.

(81) Meng, Q.; Zhu, L.; Zhang, Z. Highly Enantioselective Sequential Hydrogenation of Ethyl 2-Oxo-4-arylbut-3-enoate to Ethyl 2-Hydroxy-4-arylbutyrate. *J. Org. Chem.* **2008**, *73* (18), 7209–7212. https://doi.org/10.1021/jo801140j.

(82) Farrar-Tobar, R. A.; Dell'Acqua, A.; Tin, S.; de Vries, J. G. Metal-catalysed selective transfer hydrogenation of α,β-unsaturated carbonyl compounds to allylic alcohols. *Green Chem.* **2020**, *22* (11), 3323–3357. https://doi.org/10.1039/d0gc00855a.

(83) Meyers, A. I.; Brown, J. D. Stereoselective reduction of C=X by a chiral 1,4-dihydropyridine (NADH-MIMIC) in a self-immolative process. *Tetrahedron Lett.* **1988**, *29* (44), 5617–5620. https://doi.org/10.1016/s0040-4039(00)80827-8.

(84) Ohno, A.; Yasuma, T.; Nakamura, K.; Oka, S. NAD(P)+–NAD(P)H Models. 59. 1,2- Versus 1,4-Reduction of β,γ-Unsaturated α-Keto Ester. *Bull. Chem. Soc. Jpn.* **1986**, *59* (9), 2905–2906. https://doi.org/10.1246/bcsj.59.2905.

(85) Villegas-Torres, M. F.; Martinez-Torres, R. J.; Cázares-Körner, A.; Hailes, H.; Baganz, F.; Ward, J. Multi-step biocatalytic strategies for chiral amino alcohol synthesis. *Enzyme Microb. Technol.* **2015**, *81*, 23–30. https://doi.org/10.1016/j.enzmictec.2015.07.003.

(86) Vivat-Hannah, V.; You, D.; Rizzo, C.; Daris, J.-P.; Lapointe, P.; Zusi, F. C.; Marinier, A.; Lorenzi, M. V.; Gottardis, M. M. Synergistic cytotoxicity exhibited by combination treatment of selective retinoid ligands with taxol (Paclitaxel). *Cancer Res.* **2001**, *61*, 8703–8711.

(87) Klaholz, B. P.; Mitschler, A.; Belema, M.; Zusi, C.; Moras, D. Enantiomer discrimination illustrated by high-resolution crystal structures of the human nuclear receptor hRARγ. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97* (12), 6322–6327. https://doi.org/10.1073/pnas.97.12.6322.

(88) Chidambaram, R.; Kant, J.; Zhu, J.; Lajeunesse, J.; Sirard, P.; Ermann, P.; Schierling, P.; Lee, P.; Kronenthal, D. A Practical Synthesis of the RARγ Agonist, BMS-270394. *Org. Process Res. Dev.* **2002**, *6* (5), 632–636. https://doi.org/10.1021/op0202134.

(89) Espinosa, M.; Iborra-Torres, A.; Sanz-Marco, A.; Blay, G.; Cardona, L.; Fernández, I.; Pedro, J. R. Mg/BOX complexes as efficient catalysts for the enantioselective Michael addition of malonates to β-trifluoromethyl-α,β-unsaturated ketones and their N-tosyl imines. *Tetrahedron* **2021**, *80*, 131897. https://doi.org/10.1016/j.tet.2020.131897.

(90) Li, D.; Zhang, M.; Yang, Y.; Peng, T.; Yang, D.; Gao, W.; Wang, R. Desymmetrization Process by Mg(II)-Catalyzed Intramolecular Vinylogous Michael Reaction. *Org. Lett.* **2020**, *22* (23), 9229–9233. https://doi.org/10.1021/acs.orglett.0c03417.

(91) Bao, H.; Wu, J.; Li, H.; Wang, Z.; You, T.; Ding, K. Enantioselective Ring Opening Reaction of meso-Epoxides with Aromatic and Aliphatic Amines Catalyzed by Magnesium Complexes of BINOL Derivatives. *Eur. J. Org. Chem.* **2010**, *2010* (35), 6722–6726. https://doi.org/10.1002/ejoc.201001222.

(92) Hatano, M.; Horibe, T.; Ishihara, K. Chiral Magnesium(II) Binaphtholates as Cooperative Brønsted/Lewis Acid–Base Catalysts for the Highly Enantioselective Addition of Phosphorus Nucleophiles to α,β-Unsaturated Esters and Ketones. *Angew. Chem., Int. Ed.* **2013**, *52* (17), 4549–4553. https://doi.org/10.1002/anie.201300938.

(93) Hatano, M.; Horibe, T.; Ishihara, K. Magnesium(II)-Binaphtholate as a Practical Chiral Catalyst for the Enantioselective Direct Mannich-Type Reaction with Malonates. *Org. Lett.* **2010**, *12* (15), 3502–3505. https://doi.org/10.1021/ol101353r.

(94) Jaszczewska-Adamczak, J. A.; Mlynarski, J. Asymmetric Epoxidation of Enones Promoted by Dinuclear Magnesium Catalyst. *Adv. Synth. Catal.* **2021**, *363* (17), 4247–4255. https://doi.org/10.1002/adsc.202100482.

(95) Jaszczewska-Adamczak, J. A.; Baczewska, P.; Bujok, R.; Mlynarski, J. Magnesium-Catalyzed Asymmetric Thia-Michael Addition to α,β-Unsaturated Ketones. *Adv. Synth. Catal.* **2024**, *366* (6), 1412–1421. https://doi.org/10.1002/adsc.202301414.

(96) Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29* (5), 1189–1232. https://doi.org/10.1214/aos/1013203451.

(97) Jorner, K.; Turcani, L. *kjelljorner/morfeus: v0.7.2*; Zenodo, 2022. https://doi.org/10.5281/zenodo.7017599.

(98) Verloop, A.; Hoogenstraaten, W.; Tipker, J. Development and Application of New Steric Substituent Parameters in Drug Design. In *Drug Design*; Ariëns, E. J., Ed.; Medicinal Chemistry: A

Series of Monographs, Vol. 11; Academic Press, 1976; pp 165–207. https://doi.org/10.1016/B978-0-12-060307-7.50010-9.

(99) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods* **2015**, *71*, 58–63. https://doi.org/10.1016/j.ymeth.2014.08.005.

(100) Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *arXiv* **2020**, 2010.09885. https://doi.org/10.48550/arXiv.2010.09885.

(101) Gasteiger, J.; Marsili, M. A new model for calculating atomic charges in molecules. *Tetrahedron Lett.* **1978**, *19* (34), 3181–3184. https://doi.org/10.1016/S0040-4039(01)94977-9.

PUBLIKACJE ORYGINALNE I OŚWIADCZENIA AUTORÓW

8. Machine Learning Algorithm Guides Catalyst Choices for Magnesium-Catalyzed Asymmetric Reactions

Materiały uzupełniające dla publikacji (53 strony) są dostępne pod adresem:

http://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1002%2Fanie.202318487&file=anie202318487-sup-0001-misc_information.pdf

W skład materiałów uzupełniających wchodzą:

- Model i dane (s. 3–17) — budowa i ujednolicenie bazy danych, rozkłady nadmiaru enancjomerycznego, reprezentacje molekularne, architektura i uczenie sieci neuronowej, analiza istotności cech metodą SHAP oraz wpływ warunków reakcji na wskazania ligandów.
- Aplikacja sieciowa (s. 18–19) — opis narzędzia udostępniającego model.
- Część eksperymentalna (s. 20–23) — informacje ogólne, procedury redukcji i addycji 1,4, synteza substratów oraz charakterystyka produktów.
- Dane analiz (s. 24–51) — chromatogramy HPLC i widma NMR.
- Literatura (s. 52).

GDCh Research Article Angewandte International Edition Chemie www.angewandte.org
Check for updates
How to cite: *Angew. Chem. Int. Ed.* **2024**, *63*, e202318487
doi.org/10.1002/anie.202318487

Asymmetric Catalysis

Machine Learning Algorithm Guides Catalyst Choices for Magnesium-Catalyzed Asymmetric Reactions

Paulina Baczewska, Michał Kulczykowski, Bartosz Zambroń, Joanna Jaszczewska-Adamczak, Zbigniew Pakulski, Rafał Roszak, Bartosz A. Grzybowski, and Jacek Mlynarski**

Abstract: Organic-chemical literature encompasses large numbers of catalysts and reactions they can effect. Many of these examples are published merely to document the catalysts' scope but do not necessarily guarantee that a given catalyst is "optimal"—in terms of yield or enantiomeric excess—for a particular reaction. This paper describes a Machine Learning model that aims to improve such catalyst-reaction assignments based on the carefully curated literature data. As we show here for the case of asymmetric magnesium catalysis, this model achieves relatively high accuracy and offers out of-the-box predictions successfully validated by experiment, e.g., in synthetically demanding asymmetric reductions or Michael additions.

Introduction

Machine Learning, ML, methods are having transformative impact on fields as diverse as language models,[1] image[2] and speech recognition,[3] games of strategy,[4] protein folding,[5] and more,[6] and are also being actively studied in synthetic chemistry,[7] albeit with mixed outcomes.[8] On one hand, in synthesis planning, the all-data-driven models[9] suffer from low quality of data and lack of negative examples and, as such, are not on par with hybrid models combining expert-coded rules with elements of network theory and ML.[7c,10] On the other hand, there are other problems in which the published data appears adequate to allow predictive ML—one prominent example is the design of homogeneous catalysts for stereoselective reactions. For popular metals (Cu, Ru, Rh, Zn), the literature provides large numbers of such catalysts and reactions which they catalyze, including at least some negative examples (i.e., those with low yields or low enantiomeric excess, *ee*). Moreover, the catalysts' structures are often known, either from diffraction studies or from high-end calculations, allowing for physically sound and 3D-aware vectorization. Consequently, there have already been several success stories in ML-driven catalyst design. For instance, in their pioneering works, Sigman and co-workers created multi-descriptor models with which to predict *ee* of reactions involving chiral phosphoric acid, CPA, ligands.[11] They demonstrated the model's ability to predict enantiomeric excess across various substrates and, recently, provided a *tour de force* demonstration in which their predictions helped solve an industrial-scale synthesis problem.[12] Together with Aspuru-Guzik's group,[13] they also enumerated and, using a set of >1500 descriptors, predicted catalytic properties of some 300,000 previously unexplored organophosphorus ligands. In turn, Denmark's group proposed a ML-based method for guiding the selection of CPAs most suitable for catalyzing addition of thiols to *N*-acylimines[14] and recently, they developed a new algorithm that predicts the appropriate ligand, solvent, and base for coupling specific pairs of reactants in Pd-catalyzed C–N cross-coupling reactions.[15] List and co-workers[16] utilized binary descriptors and an initial dataset of reactions catalyzed with moderate selectivities to design improved catalysts for tetrahydropyran synthesis. Schoenebeck progressed beyond mononuclear catalysts[17] and demonstrated ML-driven design of Pd(I) dimers, which catalyze coupling of arenes with weak nucleophiles that typically do not undergo Pd^0/Pd^{II} catalytic cycle.[18]

Whereas these studies epitomize arguably the greatest challenge of catalysis-oriented ML—namely, design of novel catalysts offering improved performance in reactions of interest—there are also some "lesser" problems in which ML can be of interest and value. Here, we address one such problem in which we aim to match desired reactions against *known* catalysts that are most likely to offer the highest *ee*. The premise for this study is that ML methods should be well posed to capture combinations of reactions' features

[*] P. Baczewska, M. Kulczykowski, Dr. B. Zambroń, J. Jaszczewska-Adamczak, Dr. Z. Pakulski, Dr. R. Roszak, Prof. Dr. B. A. Grzybowski, Prof. Dr. J. Mlynarski
Institute of Organic Chemistry, Polish Academy of Sciences
Kasprzaka 44/52, 02-224 Warsaw (Poland)
E-mail: nanogrzybowski@gmail.com
jacek.mlynarski@gmail.com
Prof. Dr. B. A. Grzybowski
Center for Algorithmic and Robotized Synthesis (CARS) of Korea's Institute for Basic Science (IBS) and Department of Chemistry, Ulsan National Institute of Science and Technology 50, UNIST-gil, Eonyang-eup, Ulju-gun, Ulsan 44919 (South Korea)

© 2024 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH

15213773, 2024, 37, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202318487 by PAS/Inst Of Organic Chemistry, Wiley Online Library on [26/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(and accompanying reaction conditions) that "link" them to specific catalysts—this assumption appears correct as the accuracy of the models we develop for Mg-based catalysts is around 80 %. We validate some of the model's predictions by experiment, either in reactions for which previously used Mg-based catalysts gave lower *ee*'s or in reactions for which Mg catalysts were never used. In the future—but pending curation of high-quality datasets for ML—models like ours could be easily extended to other metals to guide more effective uses of known catalysts across a broad spectrum of reaction types.

Results and Discussion

Our studies focus on Mg-based catalysts which we find interesting for two reasons: (i) Mg itself is an earth-abundant metal, and such metals are sought as replacements for rarer and more expensive transition metals; and (ii) this system is underexplored in terms of ML and illustrates well the challenges one encounters when curating datasets with which to train robust ML models.

Dataset curation. In the context of (ii), we began by searching the Reaxys repository using keywords such as asymmetric catalysis, chiral, Mg, magnesium-catalyzed, etc. This search yielded data on almost 67,000 individual reactions but the vast majority of these records were not stereocontrolled and also lacked essential details (e.g., reaction conditions were not fully specified in ~60 % of reactions). In order to extract reactions leading to optically enriched products, we filtered off entries without *ee* assigned or catalytic conditions specified. The dataset then shrunk to only 703 reactions, making development of ML models problematic. Accordingly, we sought to augment the dataset by manual curation, starting with recent reviews by Kwit et al.[19] and Yang et al.[20] We excluded, for instance, reactions involving heterogeneous or bimetallic catalysts, multistep reactions, kinetic resolution methods, and reactions that were not fully characterized (e.g., when one yield or one *ee* value was given for a mixture of diastereomers; see Figure S1). Importantly, because we also analyzed the Supporting Information files, we were able to access reactions from the optimization domain, including some negative, low-*ee* results that rarely make their way to reaction repositories such as Reaxys. In this way, we collected additional 1054 reactions. The two sets were combined and de-duplicated by canonicalizing the SMILES of the substrates and products using the functions from MolVs library.[21] Next, we inspected the reaction conditions. Because the names of catalysts are often ambiguous or follow different naming conventions, we standardized them by preparing appropriate digital dictionaries (e.g., such that (1*R*,1′*R*,2*S*,2′*S*)-1,1′-(propane-1,3-diyl)bis(2-((2,6-diisopropylphenyl)carbamoyl)pyrrolidine-1-oxide) catalyst would be recognized as the same as $C_{37}H_{56}N_4O_4$). Furthermore, we analyzed in detail the entries in which mixtures of solvents were reported. Since solvents can dramatically alter *ee*,[22] it is important to verify whether these entries really relied on mixtures of solvents in similar proportions or some of the solvents were added only in minute quantities. To illustrate, Reaxys entry #39304844 for asymmetric [3+2]-cycloaddition between *meso*-aziridines and C3-alkylindoles catalyzed by quinine with $MgBu_2$ reported a "*n*-heptane; para-xylene" mixture. However, consulting the source publication[23] revealed that this reaction was conducted in 0.5 mL of *p*-xylene with the addition of 40 μL of $MgBu_2$ in *n*-heptane. The proper solvent to assign in this case is, obviously, *p*-xylene. After such analyses, the database of reactions and catalysts we curated encompassed 1538 reactions distributed across 139 unique ligand structures and 104 reaction types defined by the reaction cores appended with one-atom environments and key flanking groups as described in Szymkuć et al.[24] Another important factor that affects *ee* is the presence of additives[25] (e.g., 4-nitrobenzoic acid, benzyl alcohol) which can change catalyst structure, shift reaction equilibrium, or alter reaction mechanism. For ~18 % of reactions in which such additives were present, we manually unified their nomenclature and encoded their structures in canonical SMILES.

The entire dataset thus curated is posted at https://asymmetric.icho.edu.pl/download_db/. The entries provides structures of the reactants, product, chiral ligand, additives (if present), solvent or solvent mixture, temperature, and reaction *ee.* While arguably the most comprehensive compendium of stereoselective reactions affected by chiral Mg catalysts, this set is not without some obvious limitations—notably, even though it encompasses 128 examples of low-*ee* reactions (< 50 % *ee*), the overall *ee* distribution is still heavily skewed towards high *ee* values (Supporting Information, Figure S4), reflective of the bias in reporting "successful" reactions. Also, the catalysts are of unequal popularity—the most frequently encountered one is associated with 181 reactions, while the least frequent one is linked to only a single reaction.

Vectorization of reactions and conditions. Since our aim was not to alter the structures of the catalysts but to select from the known ones those that would offer the best performance for a given reaction, the catalysts were one-hot-encoded (i.e., given unique labels). By contrast, structural features were essential for the reactions whose assignment to specific catalysts we wished to learn. Accordingly, reaction substrates were represented by standard Morgan fingerprints of size 64 whereas reactions, by Differential Reaction Fingerprints (DRFP)[26] with vector size of 64 and radius 3 (for analysis of different representations, see Supporting Information). The values of temperature and *ee* were, obviously, scalars. Solvents were represented as vectors of 16 physicochemical properties taken from Stenutz' reference tables[27] (e.g., polarizability, dipole moment, dielectric constant, etc.; see Dataset available at https://asymmetric.icho.edu.pl). Additives were represented as binary Morgan fingerprints of size 64 and radius 3. If no additive was present in the reaction, zero element was used.

Neural network and feature importance. The reaction-conditions vectors thus defined served as inputs to a neural

© 2024 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH

network (which perfromed slightly better than a Random Forest classifier, see Supporting Information). The network output were are propabilties assigned the every class/catalyst and providing catalysts' ranking, from top-1 (most suitable) downwards. Here, we were most interested in the top-5 preidctions. Hyperparameters, including the activation function, layer sizes, kernel initializer, training step, and the type of layer-merging function, were optimized via the Optuna library[28] through a 300-step search designed to maximize accuracy. The final architecture, shown in Figure 1, was evaluated on our literature dataset under standard five-fold cross validation. In addition to the accuracy, the weighted average F1 score was considered to account for class imbalance. The top-1 average accuracy values—that is, the accuracies of the network predicting the catalyst really used in the literature example—ranged from 0.795 to 0.834 with the mean value of 0.811; the corresponding top-1 F-1 scores ranged from 0.756 to 0.794 with the mean of 0.772. The mean accuracies that the experimentally used catalyst would be within network's top-n predictions were top-2 = 0.854, top-3 = 0.880, top-4 = 0.885 and top-5 = 0.888.

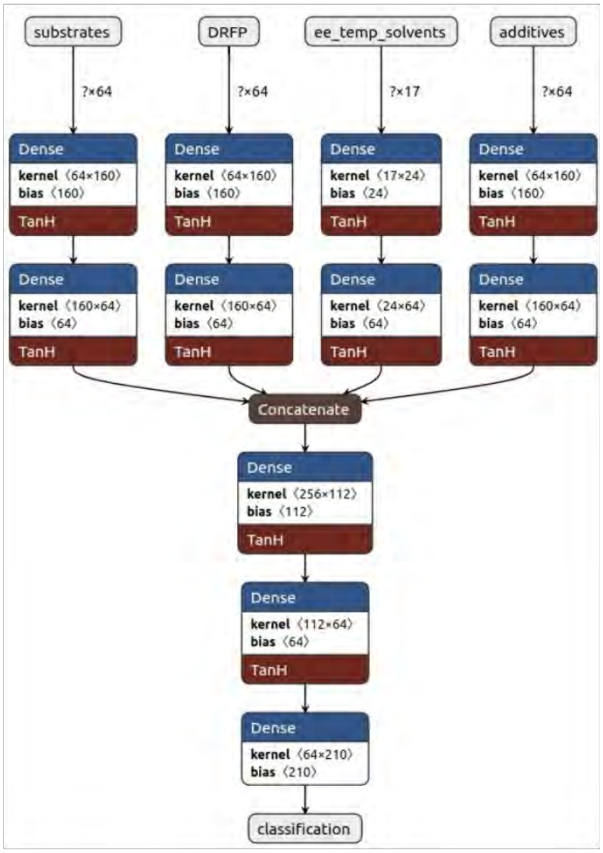


Figure 1. Architecture of the neural network after hyperparameter optimization.

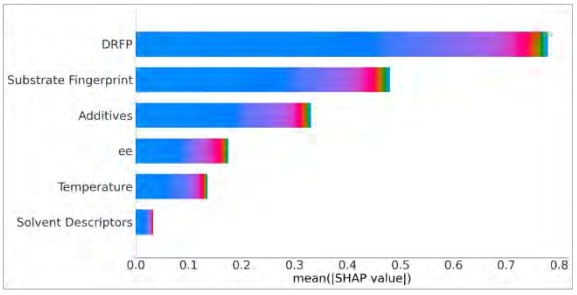


Figure 2. Feature importance quantified by SHAP values. Lengths of bars along x-axis represent mean absolute SHAP values for given features summed across all catalysts. With 139 unique ligands, each is represented by a different color. Due to the sheer number of classes, the colors appear to undergo a continuous change. The width of a particular color indicates the magnitude of the mean SHAP value for that catalyst.

With these metrics in hand, we evaluated the contributions of specific features using the SHapley Additive exPlanations (SHAP) values.[29] As illustrated in Figure 2, fingerprints for the reactions and substrates had most influence on the quality of predictions, followed by the nature of additives, *ee*, temperature and solvents. Of note, the reaction and substrate fingerprints are not redundant as each contributes independently to the prediction outcome (see Supporting Information)—this means that both the nature of the reaction "core" (accounted for by DRFP fingerprints) and structural features outside this core (accounted for by the substrate fingerprints) are important, which is congruent with common chemical knowledge.

WebApp predictor. To facilitate wider use of the model, we developed a Web Application freely available at https://asymmetric.icho.edu.pl. In the underlying NN model described above, the user must provide multiple parameters, such as the reaction Scheme in SMILES notation, temperature, solvent and, optionally, an additive. However, when planning a new experiment, this information may still be unknown. In the WebApp, users can either provide all such details or only the reaction scheme, with an option to include the additive. The model then conducts 36 predictions across six representative solvents (dichloromethane, tetrahydrofuran, toluene, 1,2-dichloroethane, diethyl ether and chlorobenzene) and six temperatures (−78 °C, −40 °C, −20 °C, 0 °C, 25 °C, 40 °C). The top-5 chiral ligands are then displayed, each with suggested temperature and solvent as well as relevant literature citation(s) specifying Mg salt used in conjunction with the specific ligand.

Experimental validation. Analyzing some of the model's predictions, we found several interesting enough to commit to experimental validation—either because they were predicted to improve the performance of other Mg-based catalysts used in the same reactions or because they expand the scope of Mg catalysis to reactions previously conducted using other metals. All validation experiments were performed at the scale of 0.1 mmol of the substrate, 1–1.5 mL of solvent and 10 % of catalyst. Unless otherwise stated, all

© 2024 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH

15213773, 2024, 37, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202318487 by PAS Inst Of Organic Chemistry, Wiley Online Library on [26/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

reactions gave conversion >95 %, and all reported yields are isolated ones.

As the first example, we considered asymmetric reduction of α-ketoesters **1a**–**3a** and leading to products **1b**, **2b** and **3b** (leftmost column in Figure 3). We chose these reactions because they were previously obtained by reduction utilizing dibutylmagnesium and chiral phosphoric acid (**L-1**) with catecholborane in mesitylene.[30] Reactions from this literature record are present in our training set, so it is not surprising that our model predicted **L-1** as the top-1 ligand for products **1b** and **3b**, and as top-2 for product **2b**. More consequentially, however, the model suggested other ligand choices of which the top-5 are listed in Figure 3. Ligands **L-2**, **L-3**, **L-5**, **L-6** were previously used by Feng[31] in various asymmetric reactions like Diels–Alder and ene reactions, α-hydroxylation of β-ketoesters and ring-opening of aziridines.[20] Feng's ligand **L-2** with ytterbium(III) triflate was used in Meerwein-Ponndorf-Verley reduction of ketoesters,[32] but since we collected information about magnesium catalysis—these reactions were unknown to our model. Ligands **L-4a**, **L-4b** were used by Gade[33] in enantioselective ketone reduction with pinacolborane. Testing ligands **L-2** and **L-3** against our reaction was feasible because they are bench stable and commercially available. Remarkably, catalyst based on $Mg(OTf)_2$ and **L-2**—predicted as top-2, top-1 and top-2 in Figure 3—offered better *ee*'s than the literature-reported **L-1**: 91 % vs. 87 % for product **1b** (with the same, quantitative yield), significantly improved 85 % vs. 66 % for **2b** (though with lower yield 83 % vs. 97 %), and 98% vs. 91% for **3b**, both at 92–93 % yield. Additionally, an appealing aspect of the ML-proposed choice is that it requires milder reaction con-

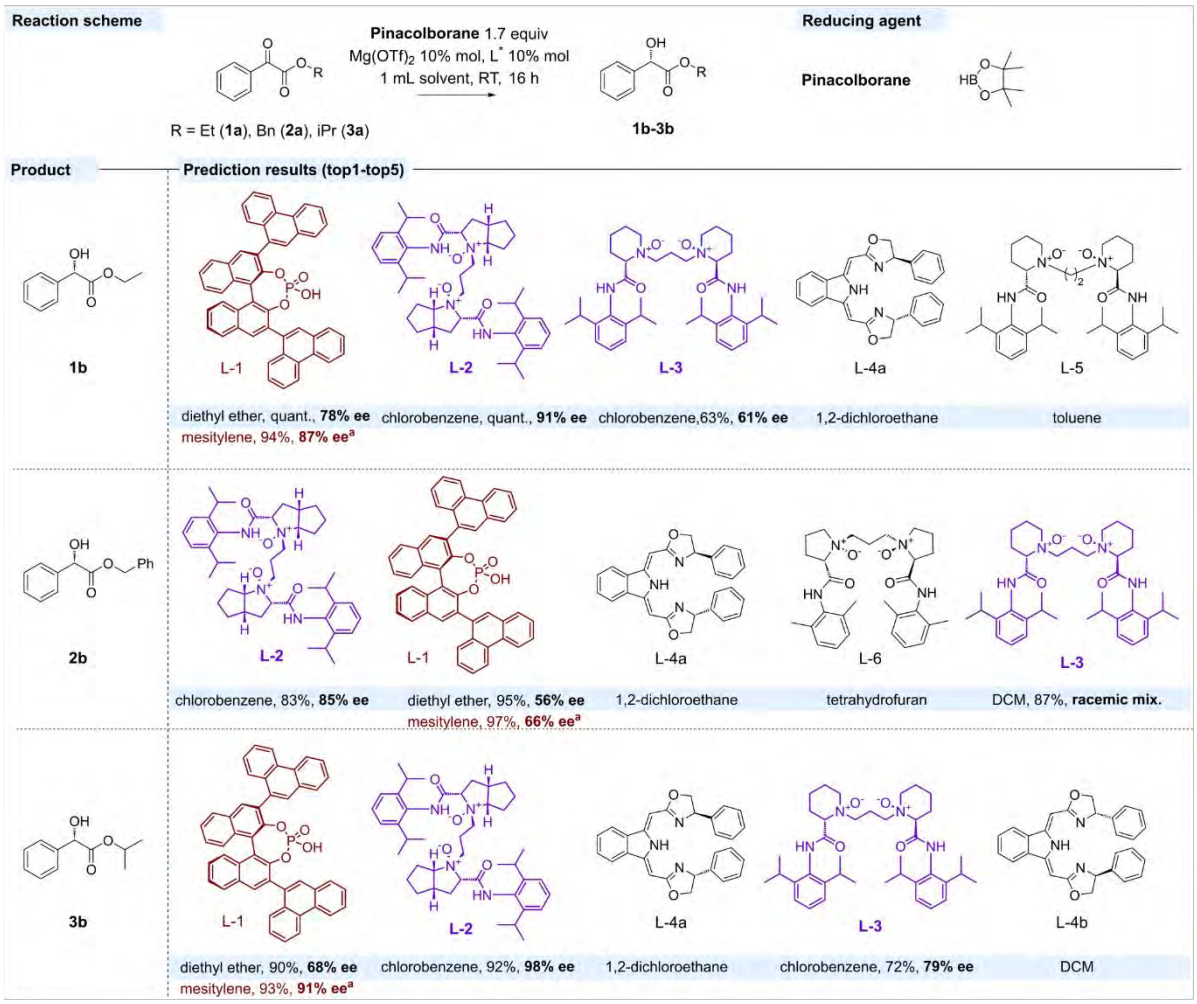


Figure 3. ML predictions and experimental validations for the reduction of α-ketoesters with HBPin (Pinacolborane). Each row represents the structure of the product, along with the top-1 to top-5 ligands predicted by our algorithm. The % isolated yields and enantiomeric excess (*ee*) values in black font are the results of experiments conducted in this study (for additional experimental information, see Supporting Information). Reaction conditions proposed by our algorithm are listed over light blue background. [a] Ligand **L-1** and the corresponding yields and *ee*'s (all colored in maroon) are taken from literature data (0.15 mmol scale, 5 % **L-1**, 5 % *n*-Bu_2Mg, catecholborane, mesitylene as solvent).[30]

© 2024 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH

15213773, 2024, 37, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202318487 by PAS Inst Of Organic Chemistry, Wiley Online Library on [26/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ditions, i.e., inexpensive and more shelf-stable $Mg(OTf)_2$ instead of more labile *n*-Bu_2Mg complexes. We also tested ligand **L-3** which was predicted with lower top-n values (top-3, top-5, top-4) than either **L-1** or **L-2**. Indeed, its performance was systematically worse, offering 61 % *ee* for product **1b**, racemic mixture for **2b**, and 79 % *ee* for **3b** (with corresponding yields 63 %, 87 %, 72 %). Finally, we also wished to test the predictions regarding the solvent itself. To this end, we re-tested this performance of ligand **L-1** in algorithm-suggested diethyl ether rather than mesitylene used in ref [30]. The outcomes were still positive, albeit with lower *ee*'s (78% vs. 87%, 56% vs. 66% and 68% vs. 91%), which is perhaps not surprising given that solvent is the least important/predictive feature of our model (see Figure 2).

Next, we carried out reactions involving ketoester **4a** featuring a conjugated double bond (top scheme in Figure 4). We chose this molecule for four reasons: (i) The motif of a chiral allylic alcohol in product **4b** is found in many natural products;[32] (ii) the reduction of the conjugated carbonyl group changes the character of the double bond from electrophilic to nucleophilic (thus enabling interesting choices for further reactivity); (iii) at the same time, substrates of this type present a challenge for asymmetric reduction due to the possibility of reducing the double bond;[34] and (iv) such reductions are typically carried out by ruthenium complexes.[35] Two literature reports from

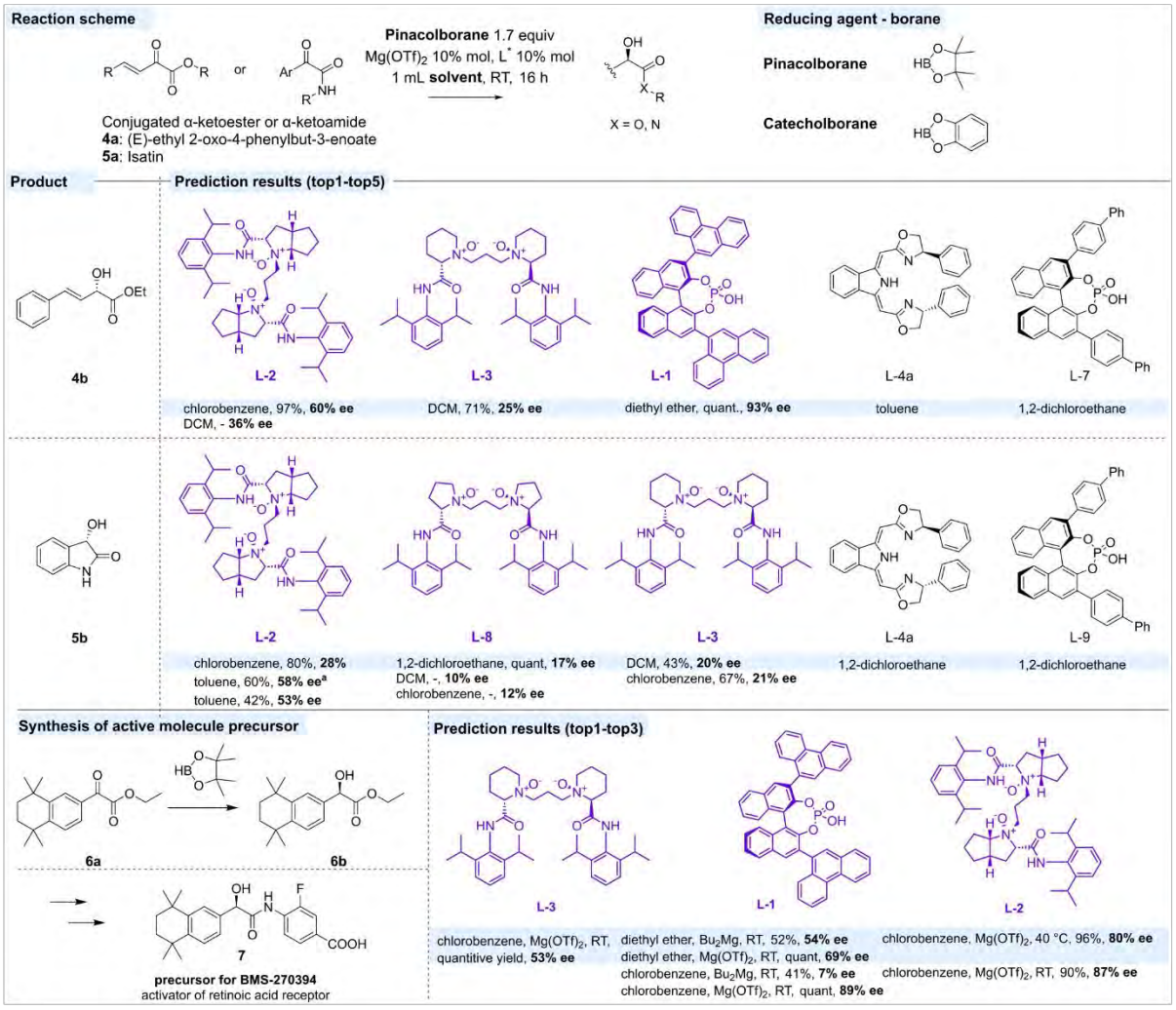


Figure 4. ML predictions and experimental validations for the magnesium-catalyzed reduction of a γ,β-unsaturated ketoester, isatin, and the precursor of the biologically active molecule BMS-270394. The % isolated yields and enantiomeric excess (*ee*) values in black font are the results of experiments conducted in this study (for additional experimental information, see Supporting Information). For three poorly performing (*ee*-wise) combinations (**4b**-**L2**-DCM; **5b**-**L8**-DCM, **5b**-**L8**-chlorobenzene), products were not isolated to determine yields. Reaction conditions on light blue background (including solvent and temperature) are based on our algorithm, while those on white background are from additional experiments we performed guided by human chemical intuition. [a] Reaction with catecholborane instead of HBPin.

© 2024 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH

15213773, 2024, 37, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202318487 by PAS Inst Of Organic Chemistry, Wiley Online Library on [26/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

15213773, 2024, 37, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202318487 by PAS/Inst Of Organic Chemistry, Wiley Online Library on [26/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

1980s[36] detail a method employing magnesium in the asymmetric reduction of substrate **4a**. However, it is important to note that this reported approach is not catalytic, as it involves the addition of 1.1 equivalents of a chiral NADH-mimic ligand and 1 equivalent of a magnesium salt to the reaction. Against this background, experiments performed with the **L-2** (top-1 prediction) and commercially available **L-3** (top-2) ligands gave good yields and *ee*'s of, respectively, 60 % and 25 % which may be of some value as a starting point for further optimization (and not overly disappointing given the cold-start nature of the problem, with no prior examples of Mg-based catalysis for this class of challenging substrates). The top-3 predicted ligand had previously been used in ketoester reduction, but only for aromatic ketoesters without an alkyl chain.[30] To our surprise, when we used ligand **L-1** with reaction algorithm-suggested conditions not typically associated with this chemistry (diethyl ether), it resulted in quantitative yield and 93 % enantiomeric excess of the product.

Our next experimental example was for isatin denoted as **5a** in Figure 4. Isatin-derived *N*-substituted compounds were previously used in asymmetric Diels–Alder or ene reactions[20] driven by Mg-catalysts based on Feng's ligands. Our algorithm suggested that these ligands (**L-2**, **L-8**, **L-3**) would also be promising (top-3) for asymmetric reduction of α-keto amides. We found this suggestion of interest because chiral 2-aminoalcohol products (e.g., **5b**) are a highly impactful motif in the design of biologically active compounds.[37]

In experiment, all ligands delivered products with some asymmetric induction but only for the top-1, **L-2** prediction was the *ee* appreciable, 50 % with solvent adjusted to toluene, and 58 % when catecholborane instead of pinacolborane was used as reducing agent. For the lower-*ee* ligands **L-8** and **L-3**, we also tried to improve the performance by changing the solvent but did not observe any significant solvent effect.

To conclude our studies on γ,β-unsaturated systems, we wished to extend beyond model substrates. In this spirit, we focused on the ketoester **6a** whose asymmetric reduction would lead to **6b** which is, in turn, a precursor of BMS-270394, **7**, an activator of retinoid acid receptor referenced in multiple patents and publications[38a,b] for its anticancer activity. In 2002, BMS improved on its own original synthesis by performing asymmetric reduction of the corresponding acid using sodium borohydride/L-tartaric acid.[38c] This protocol resulted in 75 % yield and 68 % *ee*, with enantiopurity further improved to >99 % by subsequent co-crystallization with (1*R*, 2*R*)-(−)-*N*,*N*'-dimethylcyclohexane-1,2-diamine. Against this background, our algorithm suggested Feng ligands **L-3** as top-1 and **L-2** as top-3, and CPA ligand **L-1** as top-2. The outcomes were, in all cases, positive with the highest *ee* of 80 % for **L-2**. Adjustment of solvent improved the ee value to 89% for **L-1**, while adjustment of temperature improved the ee value to 87% for **L-2**. For all reactions with $Mg(OTf)_2$, the yields were 90–99 % offering an attractive alternative to the procedure[38c] we had aimed to improve. Interestingly, the source of magnesium played a pivotal role in this example—for instance, in chlorobenzene **L-1** gave only 7 % *ee* when *n*-Bu_2Mg, was used vs. 89 % *ee* with $Mg(OTf)_2$; for reactions with dibutylmagnesium, the yields were also lower, around 50 %. Curiously, the difference was not nearly as pronounced in diethyl ether (54 % vs. 69 %), underscoring the sensitivity of magnesium-catalyzed reactions, which remains one of the main challenges in this field (and, of course, a limitation to models like ours).

The last set of validations dealt with Mg-catalyzed reactions that were published only after we have curated our dataset and were therefore not "seen" by our ML model. One such a reaction was Michael addition described by the Pedro team[39] and illustrated in Figure 5. In this case, our model was unaware of the ligand **L-10** used in Pedro's article, but suggested its close structural analog, **L-12** as top-2 choice. As the top-1 algorithm proposed the popular and commercially available BINOL (**L-11**). We tested both of these ligands under the conditions recommended by our algorithm as well as some chemically intuitive alternative conditions (e.g. non-coordinating toluene instead of diethyl ether in which **L-11** and $Mg(OTf)_2$ form a suspension). While the positive

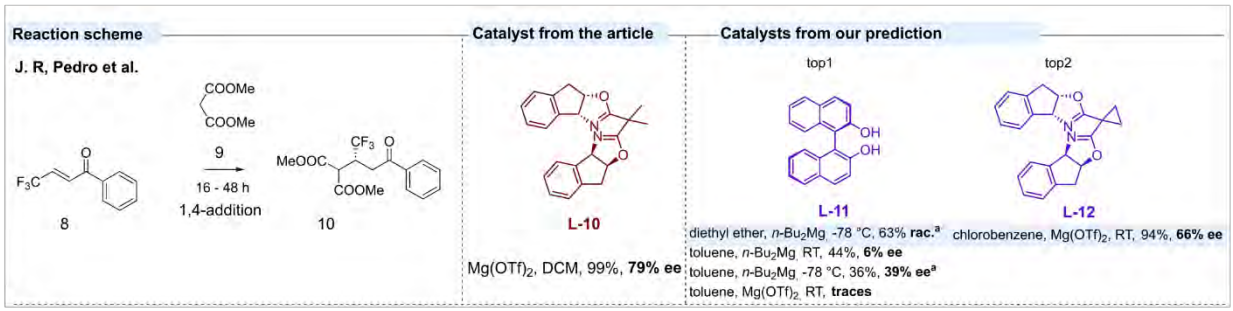


Figure 5. Magnesium complex-catalyzed 1,4-addition reaction—comparative ligand performance. Ligand colored is from recent literature; ligands we tested are in violet. The % isolated yields and enantiomeric excess (*ee*) values in black font are the results of experiments conducted in this study (for additional experimental information, see Supporting Information). Reaction conditions on light blue background (solvent and temperature) are from our algorithm, while those on white background are from additional experiments we performed guided by human chemical intuition. [a] Reactions were stirred in −78 °C for 24 h and then warm up to RT.

© 2024 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH

outcome for **L-12** (66 % *ee*) could (and maybe even "should") be expected on the basis of its close structural similarity to **L-10**, the 39 % *ee* outcome for, arguably, a very "unsophisticated" BINOL ligand was a surprise to us, suggesting that further optimization of conditions may be productive (as is often the case with BINOL ligands, see [40] or [41]). Analyzing this case in more detail, we found that our training set contained four types of BINOL-catalyzed reactions: Mannich, hetero Diels–Alder, opening of aziridine ring, and Michael addition (to unsaturated ketoesters). Of these, the specific suggestion the ML model provided was likely reflecting the examples in ref. [42,43], due to the structural similarity of the substrates.

Conclusion

In sum, the ML model we developed learned to associate substrate-reaction-condition features with specific chiral ligands that could offer improved catalytic performance. Naturally, as with any ML algorithm, ours is not without limitations. One is the relatively poor predictivity of exact reaction conditions which, however, is a pervasive problem of chemical-ML methods[8c] and, in the context of present work, reflects insufficient fundamental understanding of, e.g., solvent effects in asymmetric catalysis (especially for coordinatively labile metals like Mg), some degree of human bias (availability, habit, etc.; see [8c]), and paucity of negative reaction examples (see discussion in the Supporting Information). The other and major limitation of the model is that it will provide *some* suggestions for an "optimal" catalyst for *any* reaction the user inputs, even for reactions that cannot be catalyzed by Mg. Therefore, the algorithm's predictions should be construed as a conditional probability: Assuming that a reaction of interest *is* likely to be catalyzed by metal X, which of the existing catalysts should be used? The first part of this question will become less critical when (i) datasets incorporating examples of failed reactions become available (although harnessing negative data has, so far, proven problematic, see ref. [44] or (ii) to a lesser extent, when examples spanning multiple metals are used for training (this scenario assumes that a large fraction of organic reactions can be catalyzed by *some* metals and if so, the algorithm will satisfy this condition, proceeding to the determination of, first, the metal and then optimal catalyst choice). For this vision to materialize, however, more meticulous work is needed to curate high-quality datasets—as we showed here and before,[8c,44] automated scripting of data from incomplete or erroneous database entries does not yield ML models predictive of chemical outcomes and, instead, such data should be carefully scrutinized by expert chemists. With such datasets in hand, other ML methods and vectorization Schemes should be tested for improved performance (e.g., vectorization of ligands based on descriptors aware of three-dimensional arragments of atoms of different types[45] or of electrostatic potential maps[46]).

Supporting Information

Supporting Information provides additional theoretical and experimental details.

Author Contributions

P. B. collected and curated data, participated in model and WebApp development, performed most experiments and co-wrote the paper; M. K. curated data, led model and WebApp development and data analysis, participated in paper writing; B. Z. and J. J. A. collected data and performed experiments; Z. B. performed experiments; R. R. helped with the initial stages of model development; B. A. G. and J. M. conceived and supervised the project.

Acknowledgements

P. B., M. K., B. Z., J. A.-A., Z. P. and J. M. gratefully acknowledge support from the Polish National Science Centre (OPUS Grant NCN 2021/41/B/ST4/01617). R. R. was supported by Allchemy, Inc. Analysis of results and writing of the paper by B. A. G. was supported by the Institute for Basic Science, Korea (project code IBS-R020-D1).

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Keywords: Asymmetric catalysis · Magnesium · Machine Learning · Neural networks

[1] a) A. Graves, *ArXiv* **2013**, *abs/1308.0*; b) A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, *Adv. Neural Inf. Process. Syst.* **2017**, *2017*, 5999–6009.
[2] Y. LeCun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436–444.
[3] J. Torrejon, M. Riou, F. A. Araujo, S. Tsunegi, G. Khalsa, D. Querlioz, P. Bortolotti, V. Cros, K. Yakushiji, A. Fukushima, H. Kubota, S. Yuasa, M. D. Stiles, J. Grollier, *Nature* **2017**, *547*, 428–431.
[4] a) D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. Van Den Driessche, T. Graepel, D. Hassabis, *Nature* **2017**, *550*, 354–359; b) O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. P. Agapiou, M. Jaderberg, A. S. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T. L.

© 2024 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH

15213773, 2024, 37, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202318487 by PAS/Inst Of Organic Chemistry, Wiley Online Library on [26/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Paine, C. Gulcehre, Z. Wang, T. Pfaff, Y. Wu, R. Ring, D. Yogatama, D. Wünsch, K. McKinney, O. Smith, T. Schaul, T. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps, D. Silver, *Nature* **2019**, *575*, 350–354; c) D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, D. Hassabis, *Science* **2018**, *362*, 1140–1144.

[5] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, *Nature* **2021**, *596*, 583–589.

[6] H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Van Katwyk, A. Deac, A. Anandkumar, K. Bergen, C. P. Gomes, S. Ho, P. Kohli, J. Lasenby, J. Leskovec, T. Y. Liu, A. Manrai, D. Marks, B. Ramsundar, L. Song, J. Sun, J. Tang, P. Veličković, M. Welling, L. Zhang, C. W. Coley, Y. Bengio, M. Zitnik, *Nature* **2023**, *620*, 47–60.

[7] a) A. Bøgevig, H. J. Federsel, F. Huerta, M. G. Hutchings, H. Kraut, T. Langer, P. Löw, C. Oppawsky, T. Rein, H. Saller, *Org. Process Res. Dev.* **2015**, *19*, 357–368; b) S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk, B. A. Grzybowski, *Computer-Assisted Synthetic Planning: The End of the Beginning* **2016**; c) M. H. S. Segler, M. Preuss, M. P. Waller, *Nature* **2018**, *555*, 604–610; d) T. Klucznik, B. Mikulak-Klucznik, M. P. McCormack, H. Lima, S. Szymkuć, M. Bhowmick, K. Molga, Y. Zhou, L. Rickershauser, E. P. Gajewska, A. Toutchkine, P. Dittwald, M. P. Startek, G. J. Kirkovits, R. Roszak, A. Adamski, B. Sieredzińska, M. Mrksich, S. L. J. Trice, B. A. Grzybowski, *Chem* **2018**, *4*, 522–532; e) B. Mikulak-Klucznik, P. Gołębiowska, A. A. Bayly, O. Popik, T. Klucznik, S. Szymkuć, E. P. Gajewska, P. Dittwald, O. Staszewska-Krajewska, W. Beker, T. Badowski, K. A. Scheidt, K. Molga, J. Mlynarski, M. Mrksich, B. A. Grzybowski, *Nature* **2020**, *588*, 83–88.

[8] a) A. Tanaka, T. Kawai, T. Takabatake, N. Oka, H. Okamoto, M. Bersohn, *Tetrahedron Lett.* **2006**, *47*, 6733–6737; b) E. L. M. van Rozendaal, M. A. Ott, H. W. Scheeren, *Recl. Trav. Chim. Pays-Bas* **1994**, *113*, 297–303; c) W. Beker, R. Roszak, A. Wolos, N. H. Angello, V. Rathore, M. D. Burke, B. A. Grzybowski, *J. Am. Chem. Soc.* **2022**, *144*, 4819–4827; d) K. V. Chuang, M. J. Keiser, *Science* **2018**, *362*, 1–3.

[9] a) K. Molga, E. P. Gajewska, S. Szymkuć, B. A. Grzybowski, *React. Chem. Eng.* **2019**, *4*, 1506–1521; b) N. Fujinuma, B. DeCost, J. Hattrick-Simpers, S. E. Lofland, *Commun. Mater.* **2022**, *3*, 1–9.

[10] a) T. Badowski, E. P. Gajewska, K. Molga, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2020**, *59*, 725–730; b) A. Wołos, D. Koszelewski, R. Roszak, S. Szymkuć, M. Moskal, R. Ostaszewski, B. T. Herrera, J. M. Maier, G. Brezicki, J. Samuel, J. A. M. Lummiss, D. T. McQuade, L. Rogers, B. A. Grzybowski, *Nature* **2022**, *604*, 668–676; c) Y. Lin, R. Zhang, D. Wang, T. Cernak, *Science* **2023**, *379*, 453–457.

[11] J. P. Reid, M. S. Sigman, *Nature* **2019**, *571*, 343–348.

[12] J. J. Dotson, L. van Dijk, J. C. Timmerman, S. Grosslight, R. C. Walroth, F. Gosselin, K. Püntener, K. A. Mack, M. S. Sigman, *J. Am. Chem. Soc.* **2023**, *145*, 110–121.

[13] T. Gensch, G. dos Passos Gomes, P. Friederich, E. Peters, T. Gaudin, R. Pollice, K. Jorner, A. Nigam, M. Lindner-D'Addario, M. S. Sigman, A. Aspuru-Guzik, *J. Am. Chem. Soc.* **2022**, *144*, 1205–1217.

[14] A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow, S. E. Denmark, *Science* **2019**, *363*, eaau5631.

[15] N. I. Rinehart, R. K. Saunthwal, J. Wellauer, A. F. Zahrt, L. Schlemper, A. S. Shved, R. Bigler, S. Fantasia, S. E. Denmark, *Science* **2023**, *381*, 965–972.

[16] N. Tsuji, P. Sidorov, C. Zhu, Y. Nagata, T. Gimadiev, A. Varnek, B. List, *Angew. Chem. Int. Ed.* **2023**, *62*, e202218659.

[17] J. A. Hueffel, T. Sperger, I. Funes-Ardoiz, J. S. Ward, K. Rissanen, F. Schoenebeck, *Science* **2021**, *374*, 1134–1140.

[18] C. Fricke, T. Sperger, M. Mendel, F. Schoenebeck, *Angew. Chem. Int. Ed.* **2021**, *60*, 3355–3366.

[19] A. M. Czombik, J. Gajewy, A. Czapik, M. Kwit, *Polyhedron* **2022**, *219*, 115790.

[20] D. Yang, L. Wang, D. Li, R. Wang, *Chem* **2019**, *5*, 1108–1166.

[21] M. Swain, "MolVs," can be found under molvs.readthedocs.io/.

[22] a) H. Wynberg, B. Greijdanus, *J. Chem. Soc. Chem. Commun.* **1978**, 427–428; b) G. Li, B. Wang, D. E. Resasco, *Surf. Sci. Rep.* **2021**, *76*, 100541.

[23] L. Wang, D. Yang, F. Han, D. Li, D. Zhao, R. Wang, *Org. Lett.* **2015**, *17*, 176–179.

[24] S. Szymkuć, T. Badowski, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2021**, *60*, 26226–26232.

[25] a) E. M. Vogl, H. Gröger, M. Shibasaki, *Angew. Chem. Int. Ed.* **1999**, *38*, 1570–1577; b) L. Hong, W. Sun, D. Yang, G. Li, R. Wang, *Chem. Rev.* **2016**, *116*, 4006–4123.

[26] D. Probst, P. Schwaller, J.-L. Reymond, *Digit. Discov.* **2022**, *1*, 91–97.

[27] R. Stenutz, "Tables for Organic Chemistry," can be found under http://www.stenutz.eu/chem/.

[28] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* **2019**, 2623–2631.

[29] S. M. Lundberg, S. I. Lee, *Adv. Neural Inf. Process. Syst.* **2017**, *2017*, 4766–4775.

[30] D. Enders, B. A. Stöckel, A. Rembiak, *Chem. Commun.* **2014**, *50*, 4489–4491.

[31] a) J. Li, Y. Liao, Y. Zhang, X. Liu, L. Lin, X. Feng, *Chem. Commun.* **2014**, *50*, 6672–6674; b) C. Yin, W. Cao, L. Lin, X. Liu, X. Feng, *Adv. Synth. Catal.* **2013**, *355*, 1924–1930; c) W. Xiao, L. Ning, S. Xin, S. Dong, X. Liu, X. Feng, *Angew. Chem. Int. Ed.* **2022**, *61*.

[32] H. Qiu, J. Ren, L. Zhang, R. Song, W. Si, D. Yang, L. Wen, J. Lv, *Green Chem.* **2023**, *25*, 3948–3955.

[33] V. Vasilenko, C. K. Blasius, H. Wadepohl, L. H. Gade, *Chem. Commun.* **2020**, *56*, 1203–1206.

[34] Q. Meng, L. Zhu, Z. Zhang, *J. Org. Chem.* **2008**, *73*, 7209–7212.

[35] R. A. Farrar-Tobar, A. Dell'Acqua, S. Tin, J. G. de Vries, *Green Chem.* **2020**, *22*, 3323–3357.

[36] a) A. I. Meyers, J. D. Brown, *Tetrahedron Lett.* **1988**, *29*, 5617–5620; b) A. Ohno, T. Yasuma, K. Nakamura, S. Oka, *Bull. Chem. Soc. Jpn.* **1986**, *59*, 2905–2906.

[37] M. F. Villegas-Torres, R. J. Martinez-Torres, A. Cazeres, H. Hailes, F. Baganz, J. Ward, *Enzyme Microb. Technol.* **2015**, *81*, 23–30.

[38] a) V. Vivat-Hannah, D. You, C. Rizzo, J.-P. Daris, P. Lapointe, F. C. Zusi, A. Marinier, M. V. Lorenzi, M. M. Gottardis, *Cancer Res.* **2001**, *61*, 8703–8711; b) B. P. Klaholz, A. Mitschler, M. Belema, C. Zusi, D. Moras, *Proc. Nat. Acad. Sci.* **2000**, *97*, 6322–6327; c) R. Chidambaram, J. Kant, J. Zhu, J. Lajeunesse, P. Sirard, P. Ermann, P. Schierling, P. Lee, D. Kronenthal, *Org. Process Res. Dev.* **2002**, *6*, 632–636.

[39] M. Espinosa, A. Iborra-Torres, A. Sanz-Marco, G. Blay, L. Cardona, I. Fernández, J. R. Pedro, *Tetrahedron* **2021**, *80*.

[40] D. Li, M. Zhang, Y. Yang, T. Peng, D. Yang, W. Gao, R. Wang, *Org. Lett.* **2020**, *22*, 9229–9233.

© 2024 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH

15213773, 2024, 37, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202318487 by PAS Inst Of Organic Chemistry, Wiley Online Library on [26/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[41] H. Bao, J. Wu, H. Li, Z. Wang, T. You, K. Ding, *Eur. J. Org. Chem.* **2010**, *2010*, 6722–6726.
[42] M. Hatano, T. Horibe, K. Ishihara, *Angew. Chem. Int. Ed.* **2013**, *52*, 4549–4553.
[43] M. Hatano, T. Horibe, K. Ishihara, *Org. Lett.* **2010**, *12*, 3502–3505.
[44] F. Strieth-Kalthoff, S. Szymkuć, K. Molga, A. Aspuru-Guzik, F. Glorius, B. A. Grzybowski, *J. Am. Chem. Soc.* **2024**, *146*, 11005–11017.
[45] a) E. Bonanno, J-P. Ebejer, *Front. Pharmacol.* **2020**, *10*, 1675–1692; b) A. M. Schreyer, T. Blundell, *J. Cheminform.* **2012**, *4*, 27–38.
[46] G. Bolcato, E. Heid, J. Boström, *J. Chem. Inf. Model.* **2022**, *62*, 1388–1398.

Manuscript received: December 4, 2023
Accepted manuscript online: June 15, 2024
Version of record online: August 12, 2024

© 2024 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH

15213773, 2024, 37, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/anie.202318487 by PAS/Inst Of Organic Chemistry, Wiley Online Library on [26/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

9. Predicting Reactivity and Enantioselectivity from Small Experimental Datasets: A Case Study in Asymmetric Magnesium Catalysis

Materiały uzupełniające dla publikacji (165 stron) są dostępne do pobrania pod adresem:

https://pubs.acs.org/doi/10.1021/acscatal.6c05090

W skład materiałów uzupełniających wchodzą:

- Informacje ogólne (s. S4).
- Modele uczenia maszynowego (s. S5–S83) — przebieg prac, obliczanie deskryptorów, konfiguracja modeli, zbiory danych i wyniki wraz z analizą zakresu stosowalności.
- Część eksperymentalna (s. S84–S92) — procedury epoksydacji i addycji tio-Michaela w wariantach racemicznym i asymetrycznym oraz synteza enonów i tioli.
- Charakterystyka produktów (s. S93–S137) — epoksydy i sulfidy: dane spektroskopowe, skręcalność oraz chromatogramy HPLC na chiralnej fazie stacjonarnej.
- Literatura (s. S138).
- Widma NMR (s. S140 i dalej, do końca pliku).

ACS Catalysis

Open Access
This article is licensed under CC-BY 4.0
pubs.acs.org/acscatalysis

Research Article

Predicting Reactivity and Enantioselectivity from Small Experimental Datasets: A Case Study in Asymmetric Magnesium Catalysis

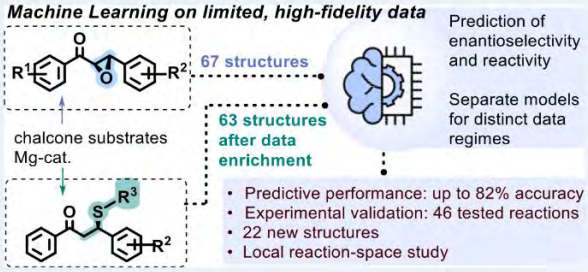
Paulina Baczewska, Damian Nowak, Joanna Jaszczewska-Adamczak, Rafał A. Bachorz, Marcin Hoffmann,* and Jacek Mlynarski*

Cite This: https://doi.org/10.1021/acscatal.6c05090

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Predicting reactivity and enantioselectivity from limited experimental data remains a central challenge in asymmetric catalysis, where reaction development is often guided by small, internally generated scope datasets rather than large screening campaigns. Here, we show that such datasets can be converted into practical decision-support tools when model development is aligned with the intrinsic limitations of low-data chemistry. Using magnesium-catalyzed chalcone epoxidation and thia-Michael addition as experimentally grounded case studies, we developed a machine-learning workflow that combines reactivity classification, where appropriate, with enantioselectivity regression trained on internally consistent laboratory data. Comparative model evaluation reveals that reactivity classification can remain robust even in highly homogeneous high-enantiomeric excess (ee) datasets, whereas reliable enantioselectivity regression requires broader selectivity distributions and careful assessment of the applicability domain. Prospective experimental validation on previously untested substrates confirms that the models capture useful structure–reactivity and structure–selectivity trends rather than merely fitting retrospective data. These results establish a realistic framework for applying machine learning to asymmetric reaction-scope development under small-data conditions, providing practical guidance for substrate prioritization and AI-assisted catalysis.



KEYWORDS: *asymmetric catalysis, machine learning, magnesium catalysts, enantioselectivity prediction, small-data modeling*

INTRODUCTION

The prediction of chemical reactivity and selectivity remains a central goal in synthetic organic chemistry.[1] Among these problems, enantioselectivity is especially difficult to forecast because it arises from small energy differences, typically 1–3 kcal mol^{-1},[2] between competing stereodetermining pathways and depends on a subtle interplay of catalyst structure, substrate properties, and reaction environment.[3] Although density functional theory (DFT) calculations of diastereomeric transition states have provided valuable mechanistic insight, they have not become a routine solution for broad selectivity prediction because of their computational cost, time requirements, and limited robustness in complex catalytic systems.

In this context, artificial intelligence (AI) and machine learning (ML) have emerged as powerful tools in synthesis, with growing impact on molecular design,[4] reaction planning,[5] and selectivity prediction (Figure 1A).[2,6] In asymmetric catalysis, ML has shown that stereochemical outcomes can in principle be learned from experimental and computed descriptors,[6a,7] sometimes with sufficient accuracy to support catalyst or substrate design.[8] Beyond outcome prediction, such approaches have been used to identify key catalyst features, quantify their influence on enantioselectivity,[4a,7d,9] and suggest effective catalytic systems.[5b,10] In a recent study, we demonstrated a proof-of-concept ML model for magnesium-catalyzed asymmetric reactions with approximately 80% predictive accuracy, supported by experimental validation.[10a]

Despite these advances, a persistent challenge in ML-guided catalysis remains the reliability and representativeness of training data, regardless of whether models employ physical-organic descriptors, DFT-derived parameters, or high-dimensional chemoinformatics features.[7c,11] Sigman and coworkers pioneered data-driven catalyst design using regression models that correlate molecular features with enantioselectivity,[6a,12] with demonstrated applications in both academic and industrial settings.[5b] Machine learning has also been applied to ligand optimization by Denmark,[4a,9a] Ackermann,[13] and List,[9b] to dinuclear catalysis by Schoenebeck,[8a] and to

Received: July 2, 2026
Revised: September 4, 2026
Accepted: September 4, 2026

ACS Publications

© XXXX The Author(s). Published by American Chemical Society

Downloaded from pubs.acs.org/acscatal/article-pdf/doi/10.1021/acscatal.6c05090/6849737/cs-2026-05090q.pdf by INSTITUTE OF ORGANIC CHEMISTRY PAS user on 26 September 2026

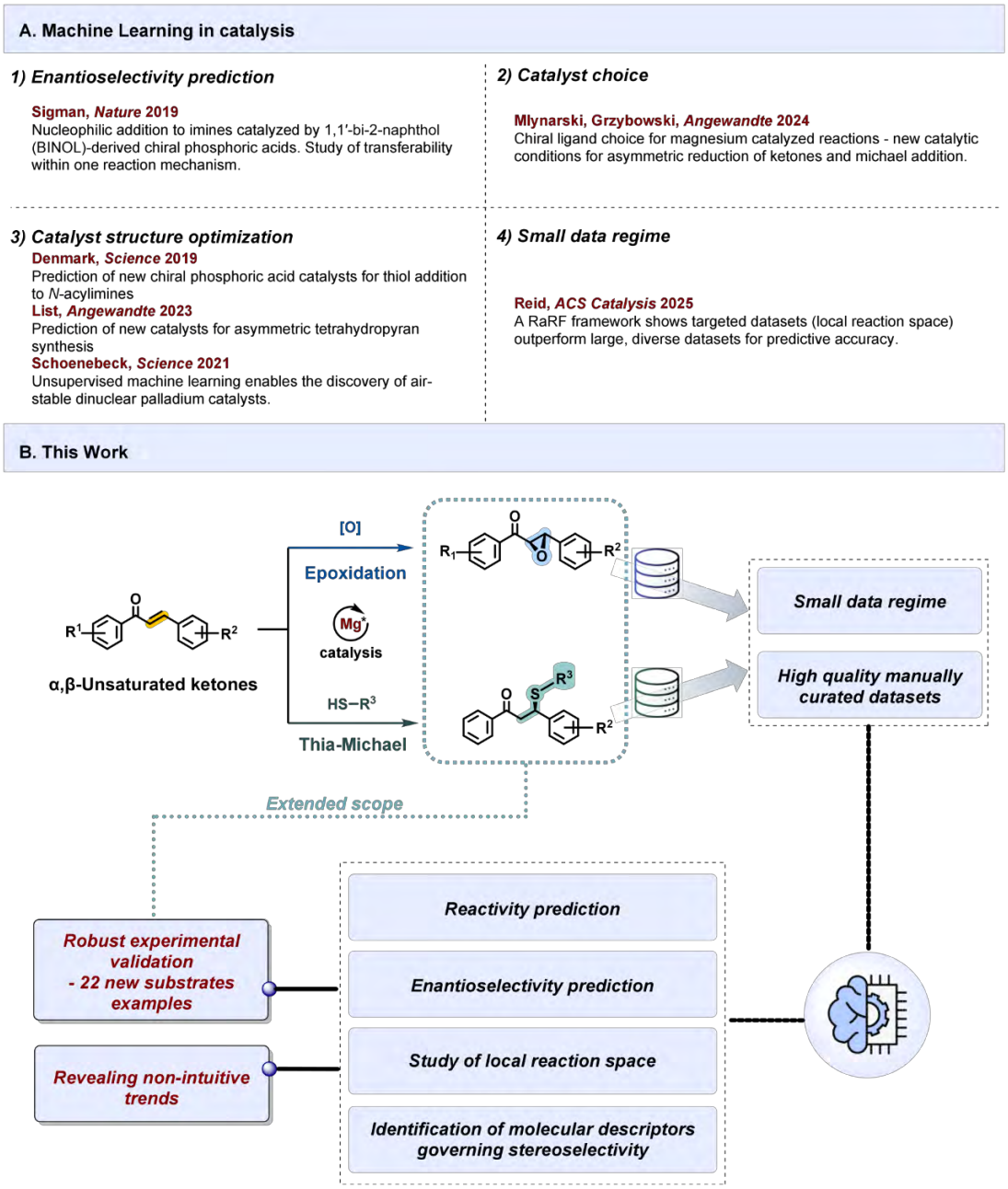


Figure 1. (A) Representative studies highlighting the development of machine learning approaches in homogeneous catalysis or asymmetric catalysis. (B) Overview of the present work, including the investigated reaction classes and the machine-learning workflow.

metal-replacement strategies by Hong and co-workers.[14] Yet, across these developments, the nature of the available data remains a central limitation: most published datasets are small, biased toward successful examples, concentrated in narrow regions of reaction space, and often affected by experimental errors.[1a,15] As a result, ML models often interpolate well within closely related substrate series, but their ability to predict reactivity or enantioselectivity beyond this local domain remains uncertain.[16]

This problem is particularly relevant in asymmetric catalysis, where scope tables frequently emphasize closely related analogues and predominantly report high yields and high enantioselectivities. While such datasets are synthetically valuable, they are less informative from a model-building perspective, especially when the goal is to distinguish between reactive and non-reactive substrates or to predict ee across structurally broader chemical space. These limitations motivate the use of small but carefully curated datasets generated under uniform

https://doi.org/10.1021/acscatal.6c05090
ACS Catal. XXXX, XXX, XXX–XXX

Downloaded from pubs.acs.org/accacs/article-pdf/doi/10.1021/acscatal.6c05090/6849737/cs-2026-05090q.pdf by INSTITUTE OF ORGANIC CHEMISTRY PAS user on 26 September 2026

experimental conditions, where data quality, internal consistency, and mechanistic relevance can compensate, at least in part, for limited size.[17] Recent studies on low-data modeling, local reaction spaces, targeted datasets, transferability, Bayesian optimization, and data selection[18] have further emphasized that data quality and distribution can be as important as algorithmic complexity.[19]

Previous studies have also shown that classification and regression can play complementary roles in data-driven reaction prediction, particularly when reaction success and selectivity are governed by partially distinct factors.[20] We therefore do not present classification-before-regression as a new machine-learning concept. Instead, we adapt this logic to small, experimentally generated asymmetric catalysis datasets, where failed reactions are chemically informative for defining accessible substrate space but cannot be treated as valid 0% ee entries in selectivity regression.

Here, we address these challenges using asymmetric magnesium catalysis as an experimentally grounded case study. We developed a machine-learning workflow (Figure 1B) centered on two complementary reaction classes: chalcone epoxidation, which provides a highly homogeneous and therefore intrinsically difficult dataset, and dinuclear Mg-promoted thia–Michael addition, in which two reaction partners are varied simultaneously, creating a more diverse prediction problem. Rather than relying on large, heterogeneous literature compilations, this approach operates on small, internally consistent datasets generated under uniform conditions in our laboratory, directly addressing the challenge of narrow reaction scopes highlighted by Kozlowski.[21]

This setup allows us to examine not only enantioselectivity prediction, but also the crucial roles of reactivity classification, descriptor choice, dataset composition, and applicability-domain analysis in defining what small-data models can realistically achieve. Rather than attempting to define a universal data-size threshold, we focus on practical substrate prioritization for everyday laboratory workflows. We demonstrate that robust reactivity classification can be achieved even under strongly biased distributions, and that useful predictive trends in enantioselectivity can be obtained from remarkably small datasets when prospective validation and domain boundaries are rigorously evaluated.

METHODS

Selection of the Reactions

Given the critical importance of data quality in predictive modeling, we based our analysis on internally generated datasets obtained under consistent experimental conditions in our laboratory. This approach ensures uniform substrate annotation, reliable determination of enantiomeric excess, and reproducible reaction conditions, minimizing sources of noise and bias commonly encountered in heterogeneous literature datasets.

Our study focuses on two representative magnesium-catalyzed asymmetric transformations: the enantioselective epoxidation of chalcones[22] and the dinuclear magnesium-promoted thia–Michael addition to α,β-unsaturated ketones (Figure 2).[23] Magnesium catalysts, owing to their earth abundance, low toxicity, and cost-effectiveness, have emerged as attractive alternatives to traditional transition-metal systems,[10a] yet their reactivity and selectivity remain difficult to predict due to the complex interplay of catalyst structure and substrate-dependent steric and electronic effects.

Under the reaction conditions developed in our laboratory, the asymmetric epoxidation of chalcones proceeds with consistently high enantiomeric excess (typically >90% ee) across the selected dataset of structurally homogeneous aryl-substituted enones. However, the origin of asymmetric induction remains unclear, with only indirect evidence suggesting the involvement of chiral magnesium peroxide species.[22,24] This combination of high selectivity and mechanistic ambiguity makes the reaction a stringent test case for data-driven approaches. At the same time, the narrow distribution of ee values introduces an inherent bias, as models trained on such datasets will naturally tend to predict high enantioselectivities for new substrates. Importantly, the inclusion of a substantial number of non-reactive substrates allowed us to formulate a reactivity classification problem, which is addressed in the following section.

In contrast, the thia–Michael addition involves two independently varied reaction partners (thiol and chalcone)[25] and spans a broader and more diverse chemical space.[23] Mechanistic insight into this transformation was also provided in our previous study. In particular, DFT calculations, supported by ^{1}H NMR experiments in the presence of the substrates, identified a low-energy, catalytically relevant Mg thiolate–chalcone assembly formed by substitution of the precatalyst *n*-butyl group with the deprotonated thiol and coordination of the enone within the catalyst sphere.[23] On this basis, the reaction is proposed to be initiated by catalyst-assisted thiol activation, followed by intracoordination-sphere attack of the activated thiol on the coordinated chalcone. The resulting intermediate then undergoes proton transfer and product release. This system therefore provides a complementary setting in which both substrate diversity and mechanistic understanding enable a more demanding assessment of model generalization and descriptor relevance. Taken together, these two reaction classes provide complementary test cases for evaluating predictive modeling under data-limited conditions: a highly homogeneous system biased toward high enantioselectivities, and a structurally diverse, two-component transformation with broader selectivity distribution. This combination allows us to systematically assess how dataset structure influences the performance of classification and regression models. Based on these datasets (Figure 2), predictive models were constructed and subsequently evaluated through experimental validation on newly synthesized substrates, as described below.

Artificial Intelligence Model

Our central objective was to evaluate whether enantioselectivity and reactivity can be predicted for structurally related but previously untested substrates using small, internally consistent datasets. To this end, we developed a machine-learning framework that systematically explores the interplay between descriptor choice, dataset composition, and model complexity under data-limited conditions (Figure 3). The workflow comprises dataset curation, descriptor generation for both substrates and products, feature selection, and systematic evaluation of classification and regression models under different validation protocols, enabling a direct comparison of factors governing predictive performance in small-data regimes.

Rather than relying on a single modeling strategy, we adopted a stepwise and systematic benchmarking approach that mirrors the workflow summarized in Figure 3. At each stage, models were constructed and evaluated to directly assess how descriptor choice, feature selection, and dataset composition influence predictive performance under small-data conditions.

In the first step, molecular representations were generated separately for both substrates and products. A broad range of descriptor families was explored, including Mordred descriptors,[26] Morgan fingerprints,[27] ChemBERTa embeddings,[28] CircuS,[9b,29] ChyLine,[9b,29] Sterimol parameters (Morfeus implementation),[30] and simple electronegativity-related descriptors derived from Gasteiger charges.[31] For details see Supporting Information file. To evaluate the contribution of individual descriptor types, multiple combinations were tested, ranging from full descriptor sets to progressively reduced representations, including models without ChemBERTa features and those based solely on physically interpretable descriptors such as Sterimol and

https://doi.org/10.1021/acscatal.6c05090
ACS Catal. XXXX, XXX, XXX–XXX

Downloaded from pubs.acs.org/accacs/article-pdf/doi/10.1021/acscatal.6c05090/6849737/cs-2026-05090q.pdf by INSTITUTE OF ORGANIC CHEMISTRY PAS user on 26 September 2026

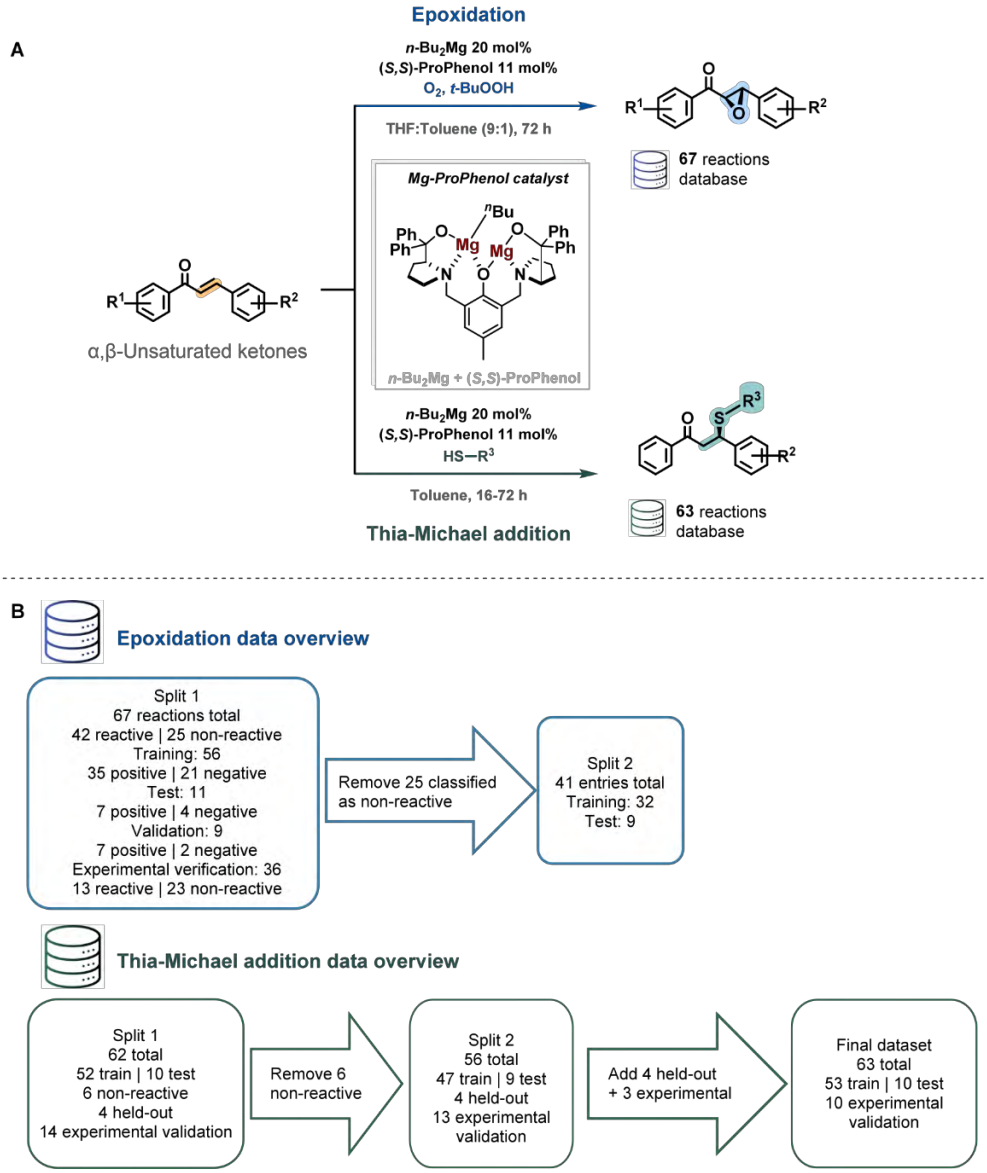


Figure 2. (A) Reaction schemes, catalytic systems, conditions, and dataset sizes used for model development. (B) Overview of the corresponding datasets.

CircuS. To systematically assess the contribution of descriptor complexity, we initially constructed models using a minimal set of physically interpretable descriptors (Sterimol parameters), and subsequently expanded the feature space to include additional descriptor families in order to evaluate whether more complex representations improve predictive performance.

For product-derived representations, hypothetical product structures were generated for all entries, including non-reactive and prospective substrates, using predefined SMARTS reaction templates. In the epoxidation dataset, the template defined the corresponding epoxy-enone framework, including the rigid epoxide core. These structures were generated from the planned substrate structures and did not contain information on the experimental outcome. The same procedure can therefore be applied prospectively before a reaction is performed. The generated structures were standardized by smart rooting at the α-carbon and converted into 3D conformers using ETKDGv3 with a fixed random seed of 42. Steric and electronic descriptors, including Sterimol parameters and Gasteiger charges, were then calculated from these geometries. Thus, product-derived descriptors encode the structure of the intended product, not whether the reaction succeeded or failed. Substrate-only and product-derived representations were benchmarked in parallel; for the epoxidation study, 840 optimized models were evaluated to compare these representations, and the corresponding results are provided in the Supporting Information and GitHub repository.

In the second step, feature selection was performed using a structured multi-stage procedure. Initial pre-selection involved removal of descriptors with more than 20% missing values, low variance (<1%),

https://doi.org/10.1021/acscatal.6c05090
ACS Catal. XXXX, XXX, XXX−XXX

Downloaded from pubs.acs.org/acscacs/article-pdf/doi/10.1021/acscatal.6c05090/6849737/cs-2026-05090q.pdf by INSTITUTE OF ORGANIC CHEMISTRY PAS user on 26 September 2026

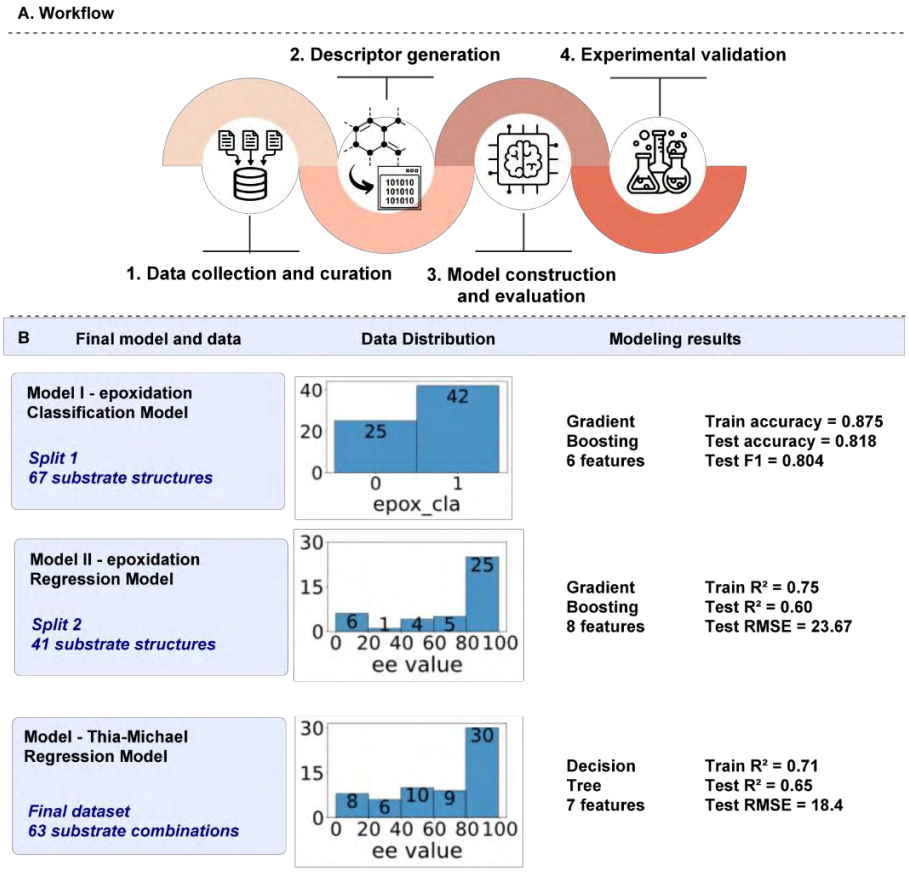


Figure 3. Overview of the machine-learning workflow and resulting models. (A) Schematic representation of the workflow, from data curation to experimental validation. (B) Summary of model types, dataset characteristics, and key predictive metrics for chalcone epoxidation and the thia–Michael reaction.

and high pairwise correlation (>95%). The reduced descriptor set was then subjected to feature selection using two complementary strategies: a fast, aggregation-based protocol (fast_aggr) and a more conservative cross-validation-based approach (cv_safe, 5-fold CV).

Consensus feature selection was used to reduce dependence on any single ranking criterion, which is particularly important in small datasets where feature importance can vary between data splits. The use of complementary ranking methods was intended to identify descriptors that remained informative across internal resampling rather than descriptors favored by one specific algorithm. The number of retained descriptors was deliberately restricted to a small range, 2–6 for classification and 3–10 for regression, to balance predictive performance, interpretability, and the risk of overfitting.

For the thia–Michael reaction, additional stratified *k*-fold procedures were applied during feature selection to assess robustness with respect to data partitioning. In the next step, extensive benchmarking, including comparison to baseline models based on mean-value prediction and repeated target shuffling (100 iterations),[32] was carried out across all descriptor combinations and feature-selection strategies. Finally, hyperparameter optimization was performed for the selected machine-learning algorithms by using the following methods: Random Forest,[33] XGBoost,[34] Gradient Boosting,[35] Logistic Regression (classification),[36] Ridge (regression),[37] Decision Tree,[38] K-Nearest Neighbors,[39] Support Vector Machines.[40] In total, 120 classification configurations (for each configuration 70 models were optimized) and 168 regression configurations (for each 98 models were optimized) models were evaluated for the chalcone epoxidation dataset, while 1008 configurations were tested for the thia–Michael reaction, corresponding to 336 configurations evaluated across three dataset variants. This systematic exploration enabled us to identify not only high-performing models, but also conditions under which predictive performance breaks down. Model selection was performed using an internal validation workflow designed to minimize information leakage. Descriptor preprocessing, imputation, scaling, feature filtering, feature selection, and hyperparameter optimization were performed using only the training data. In the cv_safe mode, feature rankings were obtained within stratified 5-fold cross-validation on the training set and averaged across complementary ranking methods, including permutation importance, recursive feature elimination, and mutual information. Hyperparameter optimization was likewise carried out by cross-validation within the training data. The external test sets were inspected only after the descriptor family, feature-selection strategy, number of features, algorithm, and hyperparameters had been fixed within the internal validation workflow. Final models were then retrained on the full training set and evaluated against the held-out test set, dummy-model baselines, and target-shuffling controls.

To avoid ambiguity in the modeling workflow, prediction targets were defined separately for reactivity classification and enantioselectivity regression. For the classification task, reactions were labeled as nonreactive only when no product formation was observed, corresponding to 0% isolated yield or no detectable product under the standard reaction conditions. A value of 0% ee was not used as a criterion for

https://doi.org/10.1021/acscatal.6c05090
ACS Catal. XXXX, XXX, XXX−XXX

Downloaded from pubs.acs.org/acscacs/article-pdf/doi/10.1021/acscatal.6c05090/6849737/cs-2026-05090q.pdf by INSTITUTE OF ORGANIC CHEMISTRY PAS user on 26 September 2026

non-reactivity. Because enantioselectivity is undefined in the absence of product formation, failed reactions were included only in the classification task and were not used as valid selectivity data for regression. Regression models were trained exclusively on reactive subsets for which product formation was experimentally confirmed and ee values were defined. The results for the best-performing reactivity classifier for the chalcone epoxidation dataset are summarized in Figure 4.

We retained %ee as the primary regression target because the objective of this study is practical substrate prioritization under routine laboratory conditions, where ee is the directly measured and operationally used output. We recognize that $\Delta\Delta G^{\ddagger}$ is often a more physically meaningful representation of enantioselectivity, particularly when reaction temperature is well defined and the goal is to connect selectivity to transition-state energy differences.[1b] By contrast, several machine-learning studies in asymmetric catalysis have modeled experimentally measured %ee directly, particularly when the goal was practical catalyst or substrate prioritization. Examples include ML models for asymmetric transfer hydrogenation,[41] asymmetric hydrogenation,[10c] asymmetric Pauson–Khand reactions,[42] and catalytic asymmetric reactions of simple alkenes,[43] all of which report or predict %ee as the primary experimental target.

In the present epoxidation dataset, however, assigning a single effective temperature to each reaction is not straightforward because the reaction involves cooling, oxygenation, subsequent warming, and extended reaction times, so product formation may occur over different parts of the temperature profile depending on substrate reactivity. Thus, conversion of all ee values into $\Delta\Delta G^{\ddagger}$ would introduce an additional assumption not directly supported by the experimental protocol. Both representations are used in the literature; here, %ee was retained to remain close to experimental decision-making while explicitly acknowledging the limitations of direct ee-based regression, especially in the high-ee regime.

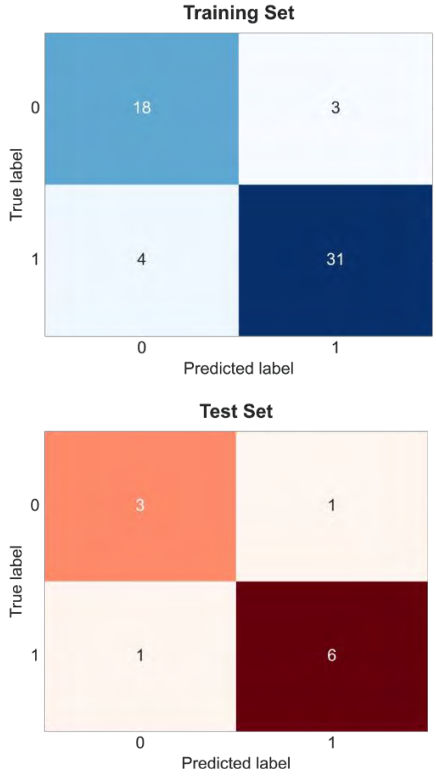


Figure 4. Confusion matrices for the best-performing reactivity classifier for chalcone epoxidation. The gradient boosting model was trained on 67 entries (56 training, 11 test) using six product-derived Sterimol/CircuS descriptors. The model achieved a balanced accuracy of 0.87 (training) and 0.80 (test), with consistent cross-validation performance (balanced accuracy 0.74 ± 0.08).

RESULTS AND DISCUSSION

For the chalcone epoxidation dataset, regression models trained on the full dataset (Figure 2. Epoxidation split 1) showed limited predictive power across all tested configurations (Figure 5A,B). Despite extensive optimization, errors remained high (RMSE ca. 30 ee points), only moderately improving over the baseline (RMSE = 41.3 ee points; dummy predictor based on the training-set mean). This behavior reflects the narrow distribution of enantiomeric excess values and confirms that regression alone is not well suited for highly homogeneous datasets (see Supporting Information for more plots).

In contrast, classification models trained on the same dataset, augmented with nonreactive substrates, showed robust and consistent performance across 120 tested configurations (see Supporting Information for tables with results). The best-performing model was a gradient boosting classifier using six product-based Sterimol and CircuS descriptors to predict reactivity as a binary outcome (0/1) for a dataset of 67 entries (56 training, 11 test; Figure 2B). On the test set, this classifier reached an accuracy of 0.818, balanced accuracy of 0.804, and macro F1 score of 0.804 (Figure 3). As shown by the confusion matrices in Figure 4, misclassification rates remained low for both classes. These results demonstrate that reactivity prediction is a well-posed problem even under strongly biased data distributions and underscore the importance of explicitly separating reactive and nonreactive regions of chemical space.

Once nonreactive cases were excluded, regression performance improved substantially. In the reduced dataset of 41 reactive entries (32 training and 9 test examples), the best model, a gradient boosting regressor based on substrate-derived descriptors and excluding ChemBERTa features, reached R^2 = 0.60, RMSE = 23.67 ee points, and MAE = 18.06 ee points on the test set (Figure 5B). Cross-validation and target-shuffling analysis further supported that this performance was robust and significantly better than noise. Together, these results show that classification followed by regression is necessary to recover meaningful structure–selectivity relationships from this small and highly skewed dataset.

Selected examples illustrating the performance of the classifier, along with the structural diversity of the corresponding substrates, are presented and discussed in the experimental validation section below.

The thia–Michael reaction provided a complementary and more demanding test case because it involves two independently varied reaction partners and a broader distribution of enantioselectivities. The initial models trained on the original thia–Michael dataset of 62 entries showed limited generalizability. In the first partition, the modeling set comprised 52 training and 10 test examples, while additional reactions were kept outside the model-development stage. Predictive performance under these conditions was unsatisfactory. Removal of nonreactive cases reduced the modeling dataset to 56 entries, but this modification alone did not provide sufficient improvement, indicating that dataset imbalance was not the only limiting factor.

https://doi.org/10.1021/acscatal.6c05090
ACS Catal. XXXX, XXX, XXX–XXX

Downloaded from pubs.acs.org/acscacs/article-pdf/doi/10.1021/acscatal.6c05090/6849737/cs-2026-05090q.pdf by INSTITUTE OF ORGANIC CHEMISTRY PAS user on 26 September 2026

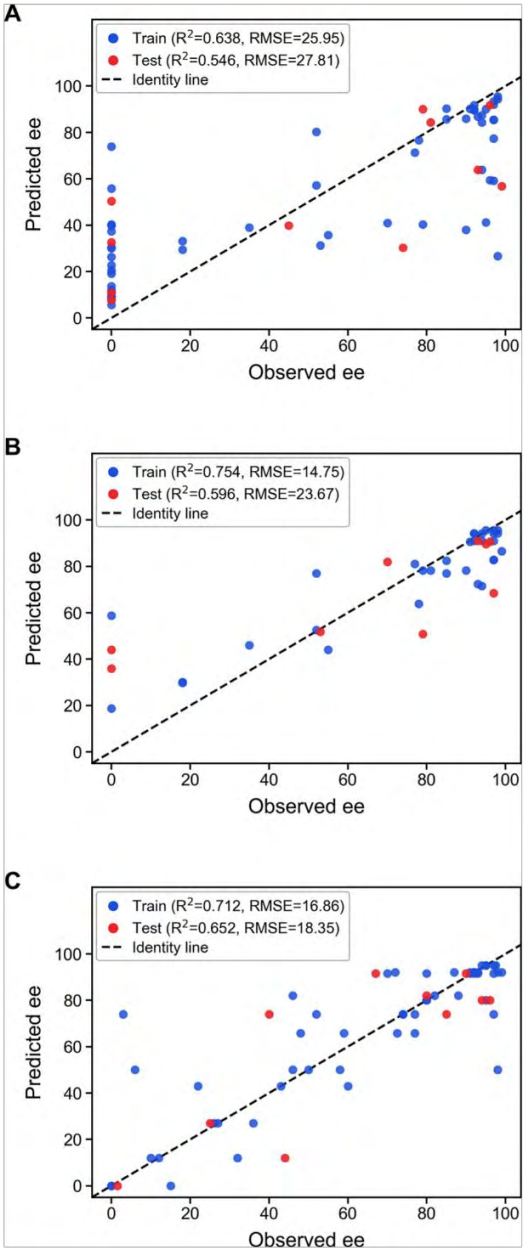


Figure 5. Predicted versus observed enantiomeric excess (ee), expressed in %, for regression models. (A) Chalcone epoxidation (full dataset). (B) Chalcone epoxidation after exclusion of non-reactive substrates. (C) Thia–Michael reaction. Model performance improves upon filtering of reactive substrates and is most consistent for the more structurally diverse thia–Michael dataset (R^2 up to 0.65 for the test set).

This initial failure was treated as an experimentally informative outcome rather than as a basis for optimizing the model around an unfavorable split. Additional reactions were therefore generated under the same experimental conditions and by the same laboratory team to increase the chemical coverage and variance of the dataset. Because some reactions that had previously been held out were subsequently incorporated into the refined dataset and the model was rebuilt, these entries are not treated as independent validation examples for the final model. They are considered part of the dataset-refinement and model-development stage.

The final thia–Michael model was developed from a refined 63-entry dataset, using a partition of 53 training and 10 test reactions. The genuinely prospective validation set consisted only of substrate combinations that remained untouched during model redevelopment and did not influence descriptor selection, feature selection, algorithm selection, hyperparameter optimization, or final model training. With this refined dataset, the final decision-tree model showed improved predictive performance, reaching $R^2 = 0.71$ and RMSE = 16.9 ee points for the training set, and $R^2 = 0.65$ and RMSE = 18.4 ee points for the test set.

For the epoxidation dataset's classification and regression tasks, we selected the Gradient Boosting model. For the thia–Michael dataset, the Decision Tree model performed best, offering a favorable balance between predictive accuracy and robustness to overfitting. SHAP analysis (Figure 6) was used as a qualitative tool to place model behavior in a chemical context rather than as definitive mechanistic proof. The top-ranked descriptors point to chemically plausible steric and electronic effects associated with model output, but their interpretation must be considered together with the small size of the datasets and the corresponding applicability-domain analysis.[44] For the epoxidation classifier, the most informative features include Sterimol parameters describing substituent size and shape, along with fragment-based descriptors sensitive to local aromatic environments. In the corresponding regression model, key descriptors reflect a combination of electronic distribution, molecular flexibility, and steric effects. For the thia–Michael reaction, the most relevant features include descriptors associated with charge distribution, molecular symmetry, and surface properties, consistent with the increased structural complexity of this system. Full details and interpretations are provided in the Supporting Information.

Taken together, these results indicate that the models capture chemically meaningful structure–selectivity relationships, while also emphasizing that predictive performance depends critically on dataset design and the integration of classification and regression within a unified framework.

Laboratory Validation (Epoxidation)

To assess the practical utility of the developed models, we performed prospective experimental validation on a set of newly synthesized substrates not included in the training datasets. Particular emphasis was placed on evaluating both reactivity classification and enantioselectivity prediction across structurally diverse examples. Epoxidations were carried out according to previously reported procedures.[22] In both cases, 11 mol % of the (*S*,*S*)-ProPhenol ligand was combined with 20 mol % of *n*-Bu_2Mg to generate the dinuclear catalyst. The catalyst was then mixed with the chalcone and cooled to −20 °C, followed by addition of TBHP, transfer to a 0 °C bath, and stirring for 72 h.

For prospective validation, 36 additional structures were first evaluated by the classifier and then tested experimentally under the standard reaction conditions. The classifier correctly identified a substantial fraction of both reactive and nonreactive substrates, but several false-positive predictions were

https://doi.org/10.1021/acscatal.6c05090
ACS Catal. XXXX, XXX, XXX–XXX

Downloaded from pubs.acs.org/acscatcs/article-pdf/doi/10.1021/acscatal.6c05090/6849375/cs-2026-05090q.pdf by INSTITUTE OF ORGANIC CHEMISTRY PAS user on 26 September 2026

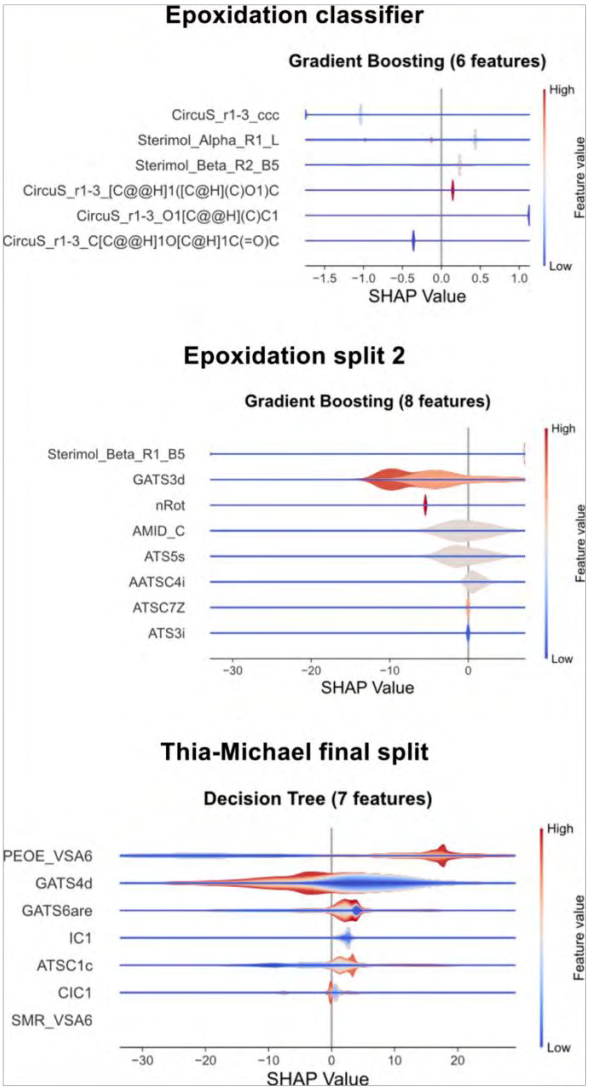


Figure 6. SHAP summary plots for the final models used for chalcone epoxidation and the thia–Michael reaction, highlighting the relative importance and directional influence of key descriptors for classification and regression tasks. SHAP values are used as qualitative indicators of model behavior and should be interpreted in the context of the small dataset size and applicability-domain analysis.

also observed, as shown in Scheme 1. Notably, substrates bearing strongly electron-withdrawing or sterically demanding substituents were consistently and correctly classified as nonreactive, in full agreement with experimental observations (see Supporting Information for representative examples). These false positives are informative because they define the limits of the classifier and indicate regions in which structural similarity to productive substrates is not sufficient to ensure reactivity. Among the substrates selected for subsequent ee regression after experimental confirmation of product formation, measurable product and defined enantiomeric excess were obtained. Thus, classification served as a useful first-stage filter, but not as a flawless predictor of productivity.

We then applied the regression model to the substrates classified as reactive in order to predict their enantiomeric excess values. As illustrated in Scheme 1, the predicted ee values reproduce the overall experimental trends for the reactive substrates. Given the limited size of the training dataset, this level of agreement is satisfactory and indicates that the model captures meaningful structure–selectivity relationships. At the same time, the predictive accuracy remains inherently limited by the narrow ee distribution in the epoxidation dataset, and the model should therefore be viewed primarily as a tool for trend prediction rather than precise quantitative forecasting. A notable exception is structure 12, which deviates strongly from the general trend: the predicted value was 73.4% ee, whereas the experimentally observed ee was only 5%. Aside from this outlier, the comparison between predicted and experimental values supports the use of the sequential classification-then-regression strategy for this reaction class.

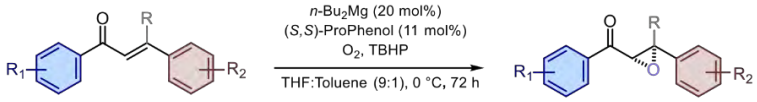
The failure of the model for compound **12**, however, is chemically informative. This substrate contains two strongly electron-withdrawing substituents, a 3-nitro group on the aryl ring adjacent to the carbonyl group and an ortho-CF_3 group on the second aryl ring. Applicability-domain analysis placed this substrate outside the reliable descriptor space of the regression model. This is consistent with the composition of the training set, which did not contain a directly analogous substrate combining such a strongly electron-deficient substitution pattern with a meta-substituted aryl ring on the carbonyl side. The model was therefore required to extrapolate into a poorly represented region of chemical space. Experimentally, compound **12** also showed problematic behavior under the standard reaction conditions, including low selectivity and reduced practical performance, consistent with the strongly electron-poor character of the substrate and its limited solubility. The solvent system was not modified because maintaining identical reaction conditions was essential for comparison across the validation set. This example illustrates the value of the applicability-domain analysis as a warning indicator for extrapolative predictions.

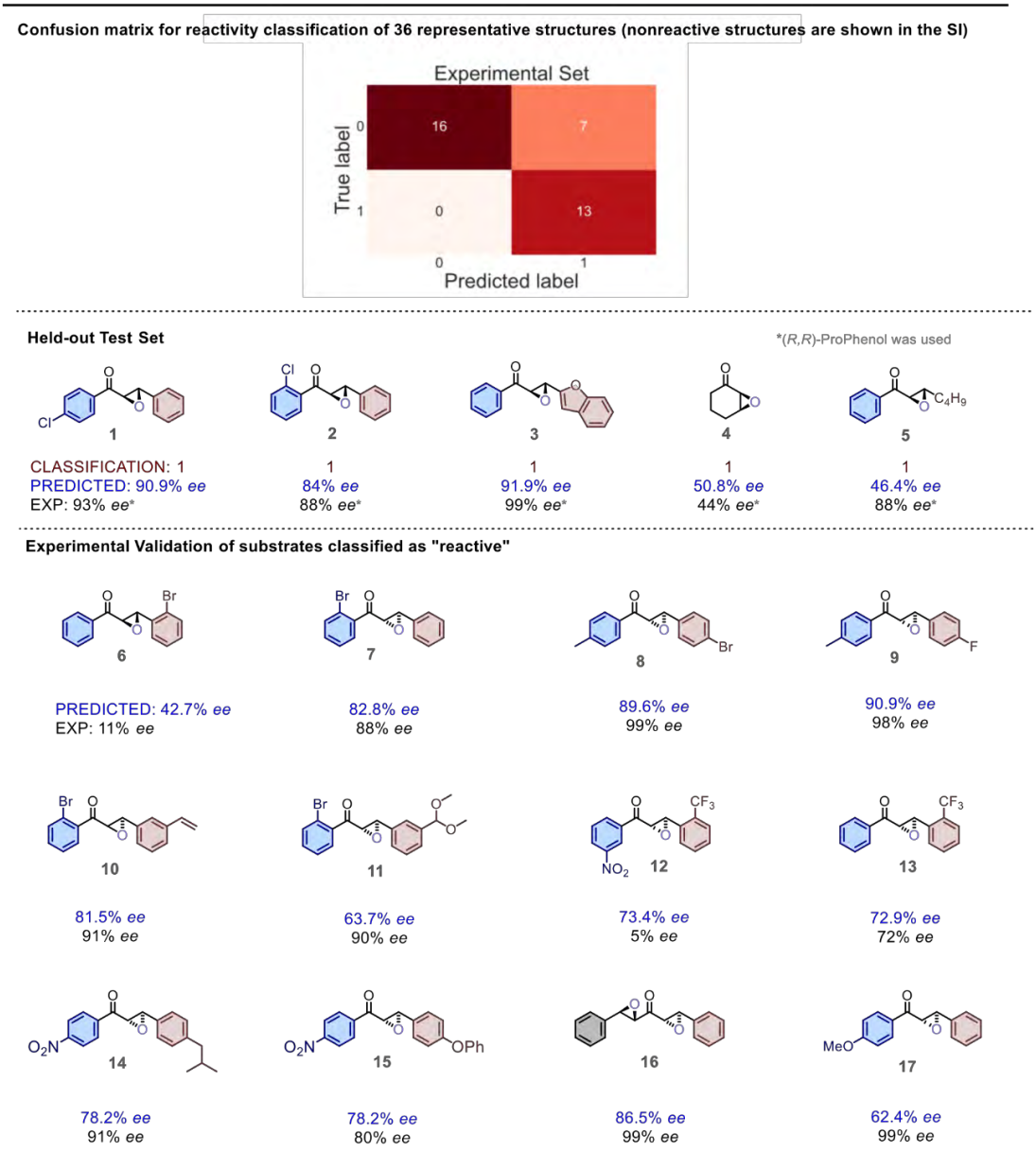
To place these validation results in context, applicability-domain analysis was used to distinguish predictions corresponding to interpolation within the training descriptor space from extrapolative predictions. For each final model, the selected descriptors were projected into a reduced PCA space, and a convex hull was constructed around the training-set compounds. Validation substrates located inside or close to this hull were treated as interpolative or near-interpolative, whereas substrates outside the hull were considered extrapolative. This analysis provides essential context for interpreting both successful and failed predictions. In particular, the strongly mispredicted epoxidation substrate **12** was located outside the reliable applicability domain, consistent with its unusual combination of strongly electron-withdrawing substituents and the absence of directly analogous examples in the training set. The applicability-domain analysis therefore supports the interpretation of the models as trend-prediction and prioritization tools within defined chemical space, rather than universal predictors of absolute ee values. The corresponding PCA/convex-hull maps are provided in the Supporting Information (Figures S36 and S37); the compound numbering used there is given in Figure S35, panel (A) for the external validation set and panel (B) for the prospective experimental set.

https://doi.org/10.1021/acscatal.6c05090
ACS Catal. XXXX, XXX, XXX–XXX

Downloaded from pubs.acs.org/accacs/article-pdf/doi/10.1021/acscatal.6c05090/6849737/cs-2026-05090q.pdf by INSTITUTE OF ORGANIC CHEMISTRY PAS user on 26 September 2026

Scheme 1. Magnesium-Catalyzed Epoxidation of Chalcones[a]





[a]Reaction scheme and conditions, reactivity classification for 36 prospectively validated substrates, and experimental validation of substrates predicted as reactive, including comparison of predicted and observed enantiomeric excess values.

Laboratory Validation (Thia-Michael)

In contrast, the thia–Michael reaction provides a more demanding validation scenario, as it involves two independently varied reaction partners and a broader distribution of enantioselectivities. Again, the thia–Michael reaction of chalcones and thiols was performed according to the previously published procedure.[23] Catalyst generation followed the same protocol as for the epoxidation process, after which the catalyst was combined with the chalcone, and the temperature was

https://doi.org/10.1021/acscatal.6c05090
ACS Catal. XXXX, XXX, XXX−XXX

Downloaded from pubs.acs.org/accacs/article-pdf/doi/10.1021/acscatal.6c05090/6849737/cs-2026-05090q.pdf by INSTITUTE OF ORGANIC CHEMISTRY PAS user on 26 September 2026

lowered to −78 °C. Upon cooling, the thiol solution was added via a syringe pump over 30 min. The reaction mixture was stirred at −78 °C for 16–48 h (see Supporting Information for details and yields).

Validation results for the refined thia–Michael dataset are shown in Scheme 2. In contrast to the chalcone epoxidation dataset, this reaction space supports a quantitatively useful regression model for enantioselectivity. Prospective validation shows that a model built exclusively from 2D Mordred descriptors retains predictive value for previously untested enone/thiol combinations. For the 10 prospective examples examined here, the model gave an MAE of 9.4 ee points, an RMSE of 14.04 ee points, and $R^2 = 0.48$, with several predictions in close agreement with experiment (e.g., 65.8/66, 92/92, 80/79, 91.5/88, and 65.8/64 [% ee]). At the same time, larger deviations for selected cases, including 50/27 and 74/43% ee, indicate that the model should be viewed primarily as a tool for ranking and prioritizing substrate combinations rather than as a uniformly accurate predictor of absolute ee values.

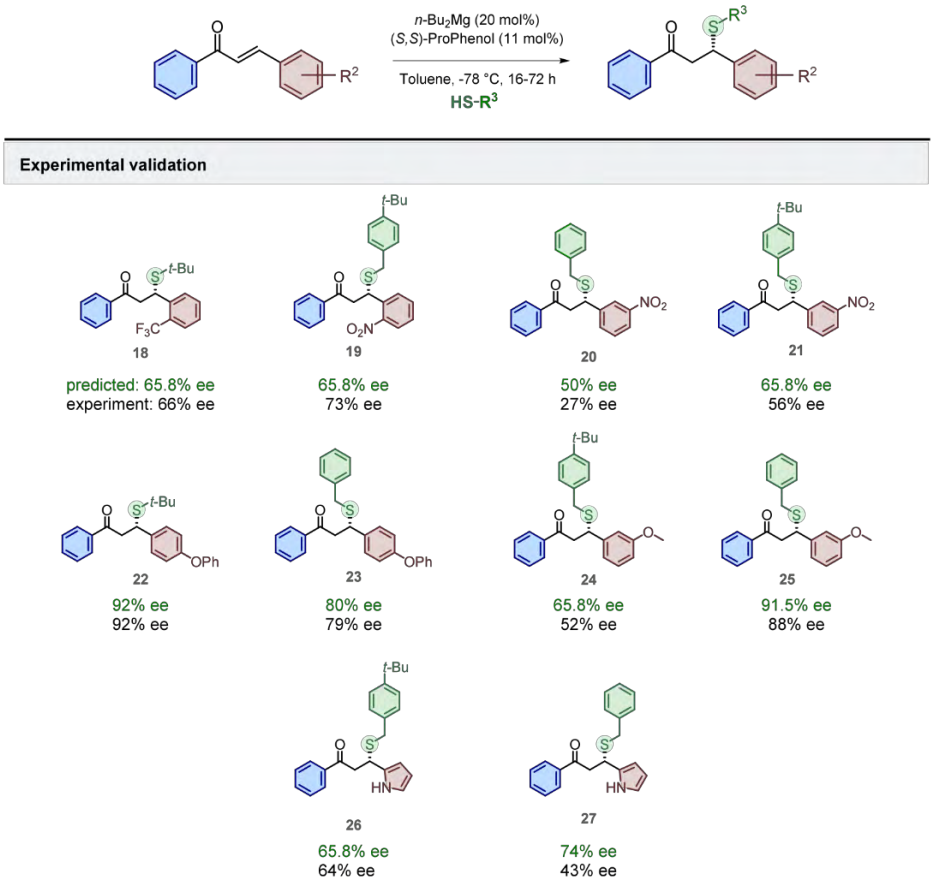
The repeated predicted values observed for several validation substrates arise from the coarse structure of the decision-tree regressor, in which chemically related substrates can fall into the same terminal node and therefore receive the same predicted ee value. For this reason, the model should be evaluated not only by absolute-error metrics but also by its ability to rank substrate combinations. Ranking-based statistics, including Spearman and Kendall correlation coefficients, are provided in the Supporting Information and support the use of the model as a substrate-prioritization tool.

Because all retained features are 2D descriptors, the predictive performance of the thia–Michael model indicates that enantioselectivity in this dataset is associated, at least in part, with topological complexity, medium-range electronic patterning, and charge- or surface-related molecular features encoded at the graph level. At the same time, the observed errors define the limits of this representation. The model should therefore be interpreted not as a mechanistic description of stereocontrol, but as an experimentally validated surrogate for structure–-selectivity relationships within the present reaction space.

CONCLUSIONS

This study demonstrates that predictive modeling in asymmetric catalysis under data-limited conditions is feasible, but strongly dependent on dataset architecture, response distribution, and the definition of the modeling task. Using Mg-catalyzed

Scheme 2. Thia–Michael Reaction[a]



[a]Reaction scheme and experimental validation of enantioselectivity prediction, with comparison of predicted and observed enantiomeric excess values. Predicted ee values are shown rounded to the precision used in the figure; the reported MAE, RMSE and R^2 were calculated from the unrounded model outputs.

https://doi.org/10.1021/acscatal.6c05090
ACS Catal. XXXX, XXX, XXX–XXX

Downloaded from pubs.acs.org/acscatal/article-pdf/doi/10.1021/acscatal.6c05090/6849737/cs-2026-05090q.pdf by INSTITUTE OF ORGANIC CHEMISTRY PAS user on 26 September 2026

asymmetric epoxidation and thia–Michael addition as experimentally grounded case studies, we show that reactivity classification and enantioselectivity regression should not be treated as equivalent prediction problems. Classification provides a robust entry point for identifying accessible regions of chemical space, whereas meaningful regression requires broader selectivity distributions, sufficient representation of low- and medium-performing examples, and careful assessment of model applicability.

For the relatively homogeneous chalcone epoxidation dataset, reactivity classification provided the most reliable and practically useful outcome, while direct regression of enantiomeric excess was intrinsically limited by the narrow high-ee distribution and sensitivity to dataset composition. In contrast, the broader thia–Michael dataset enabled construction of a regression model with measurable predictive value. Its main utility, however, lies in ranking and prioritizing substrates rather than in predicting absolute ee values with high precision. Prospective experimental validation confirmed that the models capture useful structure–reactivity and structure–selectivity trends, while also illustrating that small changes in dataset composition can have a disproportionately large effect on model performance in the low-*N* regime. Equally important, the observed failure cases, such as the strongly mispredicted epoxidation substrate **12**, show that applicability-domain analysis is essential for identifying when a small-data model is being asked to extrapolate beyond the chemical space represented in training.

More broadly, our results indicate that the success of machine learning in asymmetric catalysis is governed less by algorithmic complexity than by chemically informed dataset design, transparent validation, and explicit recognition of applicability limits. Conventional reaction-scope studies are typically optimized to demonstrate synthetic generality, not to generate balanced training data. As a result, highly successful but narrowly distributed scope tables may contain insufficient variance for reliable regression, even when they are chemically meaningful. Conversely, strategically including negative, low-selectivity, and borderline cases can substantially increase the value of small datasets for predictive modeling.

Taken together, this work establishes a practical small-data workflow that combines reactivity classification, enantioselectivity regression, applicability-domain analysis, and prospective laboratory validation. Rather than presenting machine learning as a replacement for experimentation, the approach defines a realistic role for AI in asymmetric catalysis: guiding substrate prioritization, identifying promising regions of chemical space, and helping chemists decide which experiments are most informative to perform next.

■ ASSOCIATED CONTENT

Data Availability Statement

All datasets, split assignments, random seeds, descriptor-generation settings, and Python scripts used for model construction and validation are publicly available in the GitHub repository asymmetric-catalysis-qspr-ml. The repository contains the input structures, target values, descriptor-generation workflow, feature-selection settings, model-training scripts, and validation outputs required to reproduce the computational analysis reported in this work.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acscatal.6c05090.

Supporting Information provides additional theoretical and experimental details, including, synthetic procedures, compound characterization, and spectra for the newly prepared compounds; The following files are available free of charge (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Marcin Hoffmann – *Adam Mickiewicz University, Faculty of Chemistry, Uniwersytetu Poznańskiego 8, Poznań 61-614, Poland;* Email: marcin.hoffman@amu.edu.pl

Jacek Mlynarski – *Polish Academy of Sciences, Institute of Organic Chemistry, Kasprzaka 44/52, Warsaw 01-224, Poland;* orcid.org/0000-0002-1794-306X; Email: jacek.mlynarski@gmail.com

Authors

Paulina Baczewska – *Polish Academy of Sciences, Institute of Organic Chemistry, Kasprzaka 44/52, Warsaw 01-224, Poland*

Damian Nowak – *Adam Mickiewicz University, Faculty of Chemistry, Uniwersytetu Poznańskiego 8, Poznań 61-614, Poland;* orcid.org/0000-0003-3739-3306

Joanna Jaszczewska-Adamczak – *Polish Academy of Sciences, Institute of Organic Chemistry, Kasprzaka 44/52, Warsaw 01-224, Poland*

Rafał A. Bachorz – *Institute of Medical Biology, Polish Academy of Sciences, Łódź 93-232, Poland; Institute of Computing Science, Poznan University of Technology, Poznań 60-965, Poland*

Complete contact information is available at: https://pubs.acs.org/doi/10.1021/acscatal.6c05090

Author Contributions

All authors have given approval to the final version of the manuscript.

Funding

Polish National Science Centre (NCN 2021/41/B/ST4/01617).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

J.M. and PB gratefully acknowledge support from the Polish National Science Centre (OPUS Grant NCN 2021/41/B/ST4/01617).

■ REFERENCES

(1) (a) Tan, Z.; Yang, Q.; Luo, S. AI Molecular Catalysis: Where Are We Now? *Org. Chem. Front.* **2025**, *12* (8), 2759–2776. (b) Ruth, M.; Gensch, T.; Schreiner, P. R. Contrasting Historical and Physical Perspectives in Asymmetric Catalysis: ΔΔG≠ versus Enantiomeric Excess. *Angew. Chem., Int. Ed.* **2024**, *63* (48), No. e202410308.

(2) (a) Chen, G. M.; Ye, Z. H.; Li, Z. M.; Zhang, J. L. Universal Descriptors of Quasi Transition States for Small-Data-Driven Asymmetric Catalysis Prediction in Machine Learning Model. *Cell Rep. Phys. Sci.* **2024**, *5* (7), No. 102043. (b) van Dijk, L.; Haas, B. C.; Lim, N.-K.; Clagg, K.; Dotson, J. J.; Treacy, S. M.; Piechowicz, K. A.; Roytman, V. A.; Zhang, H.; Toste, F. D.; Miller,

Downloaded from pubs.acs.org/acscacs/article-pdf/doi/10.1021/acscatal.6c05090/6849737/cs-2026-05090q.pdf by INSTITUTE OF ORGANIC CHEMISTRY PAS user on 26 September 2026

S. J.; Gosselin, F.; Sigman, M. S. Data Science-Enabled Palladium-Catalyzed Enantioselective Aryl-Carbonylation of Sulfonimidamides. *J. Am. Chem. Soc.* **2023**, *145* (38), 20959–20967.
(3) (a) Gallarati, S.; Fabregat, R.; Laplaza, R.; Bhattacharjee, S.; Wodrich, M. D.; Corminboeuf, C. Reaction-Based Machine Learning Representations for Predicting the Enantioselectivity of Organocatalysts. *Chem. Sci.* **2021**, *12* (20), 6879–6889. (b) Gallarati, S.; Bucci, E. M.; Doyle, A. G.; Sigman, M. S. Transferable Enantioselectivity Models from Sparse Data. *Nature* **2026**, *651* (8106), 637–646.
(4) (a) 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* **2019**, *363* (6424), No. eaau5631. (b) Poree, C.; Schoenebeck, F. A Holy Grail in Chemistry: Computational Catalyst Design: Feasible or Fiction? *Acc. Chem. Res.* **2017**, *50* (3), 605–608.
(5) (a) Mikulak-Klucznik, B.; Gołębiowska, P.; Bayly, A. A.; Popik, O.; Klucznik, T.; Szymkuć, S.; Gajewska, E. P.; Dittwald, P.; Staszewska-Krajewska, O.; Beker, W.; Badowski, T.; Scheidt, K. A.; Molga, K.; Mlynarski, J.; Mrksich, M.; Grzybowski, B. A. Computational Planning of the Synthesis of Complex Natural Products. *Nature* **2020**, *588* (7836), 83–88. (b) Dotson, J. J.; van Dijk, L.; Timmerman, J. C.; Grosslight, S.; Walroth, R. C.; Gosselin, F.; Püntener, K.; Mack, K. A.; Sigman, M. S. Data-Driven Multi-Objective Optimization Tactics for Catalytic Asymmetric Reactions Using Bisphosphine Ligands. *J. Am. Chem. Soc.* **2023**, *145* (1), 110–121. (c) Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem., Int. Ed.* **2016**, *55* (20), 5904–5937.
(6) (a) Reid, J. P.; Sigman, M. S. Holistic Prediction of Enantioselectivity in Asymmetric Catalysis. *Nature* **2019**, *571* (7765), 343–348. (b) Karl, T. M.; Bouayad-Gervais, S.; Hueffel, J. A.; Sperger, T.; Wellig, S.; Kaldas, S. J.; Dabranskaya, U.; Ward, J. S.; Rissanen, K.; Tizzard, G. J.; Schoenebeck, F. Machine Learning-Guided Development of Trialkylphosphine $Ni^{(I)}$ Dimers and Applications in Site-Selective Catalysis. *J. Am. Chem. Soc.* **2023**, *145* (28), 15414–15424.
(7) (a) Betinol, I. O.; Kuang, Y.; Lai, J.; Yousofi, C.; Reid, J. P. Machine Learning Enables a Top-Down Approach to Mechanistic Elucidation in Asymmetric Catalysis. *ACS Catal.* **2025**, *15* (11), 8799–8810. (b) Crawford, J. M.; Kingston, C.; Toste, F. D.; Sigman, M. S. Data Science Meets Physical Organic Chemistry. *Acc. Chem. Res.* **2021**, *54* (16), 3136–3148. (c) Cammarota, R. C.; Liu, W.; Bacsa, J.; Davies, H. M. L.; Sigman, M. S. Mechanistically Guided Workflow for Relating Complex Reactive Site Topologies to Catalyst Performance in C-H Functionalization Reactions. *J. Am. Chem. Soc.* **2022**, *144* (4), 1881–1898. (d) Ocampo, B. E.; Altundas, B.; Bock, M. J.; Feiz, S.; Denmark, S. E. Data-Driven Prediction of Enantioselectivity for the Sharpless Asymmetric Dihydroxylation: Model Development and Experimental Validation. *ACS Cent. Sci.* **2025**, *11* (9), 1640–1650.
(8) (a) Hueffel, J. A.; Sperger, T.; Funes-Ardoiz, I.; Ward, J. S.; Rissanen, K.; Schoenebeck, F. Accelerated Dinuclear Palladium Catalyst Identification through Unsupervised Machine Learning. *Science* **2021**, *374* (6571), 1134–1140. (b) Beker, W.; Roszak, R.; Wolos, A.; Angello, N. H.; Rathore, V.; Burke, M. D.; Grzybowski, B. A. Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki-Miyaura Coupling. *J. Am. Chem. Soc.* **2022**, *144* (11), 4819–4827.
(9) (a) Rinehart, N. I.; Saunthwal, R. K.; Wellauer, J.; Zahrt, A. F.; Schlemper, L.; Shved, A. S.; Bigler, R.; Fantasia, S.; Denmark, S. E. A Machine-Learning Tool to Predict Substrate-Adaptive Conditions for Pd-Catalyzed C–N Couplings. *Science* **2023**, *381* (6661), 965–972. (b) Tsuji, N.; Sidorov, P.; Zhu, C.; Nagata, Y.; Gimadiev, T.; Varnek, A.; List, B. Predicting Highly Enantioselective Catalysts Using Tunable Fragment Descriptors. *Angew. Chem., Int. Ed.* **2023**, *62* (11), No. e202218659.
(10) (a) Baczewska, P.; Kulczykowski, M.; Zambroń, B.; Jaszczewska-Adamczak, J.; Pakulski, Z.; Roszak, R.; Grzybowski, B. A.; Mlynarski, J. Machine Learning Algorithm Guides Catalyst Choices for Magnesium-Catalyzed Asymmetric Reactions. *Angew. Chem., Int. Ed.* **2024**, *63* (37), No. e202318487. (b) Tang, M.-J.; Zhang, T.; Huang, Q.; Li, S.; Liu, R.; Li, H.; Chen, X.; Dong, S.; Liu, X.; Feng, X.; Hong, X. Data-Driven Modeling of N,N′-Dioxide/Metal-Catalyzed Asymmetric Michael Additions. *Angew. Chem., Int. Ed.* **2026**, *65* (12), No. e202518560. (c) Singh, S.; Pareek, M.; Changotra, A.; Banerjee, S.; 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. U. S. A.* **2020**, *117* (3), 1339–1345.
(11) Morán-González, L.; Burnage, A. L.; Nova, A.; Balcells, D. AI Approaches to Homogeneous Catalysis with Transition Metal Complexes. *ACS Catal.* **2025**, *15* (11), 9089–9105.
(12) (a) Gensch, T.; don Passos Gomes, G.; Friederich, P.; Peters, E.; Gaudin, T.; Pollice, R.; Jorner, K.; Nigam, A.; Lindner-D'Addario, M.; Sigman, M. S.; Aspuru-Guzik, A. A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis. *J. Am. Chem. Soc.* **2022**, *144* (3), 1205–1217. (b) Akana, M. E.; Tcyrulnikov, S.; Akana-Schneider, B. D.; Reyes, G. P.; Monfette, S.; Sigman, M. S.; Hansen, E. C.; Weix, D. J. Computational Methods Enable the Prediction of Improved Catalysts for Nickel-Catalyzed Cross-Electrophile Coupling. *J. Am. Chem. Soc.* **2024**, *146* (5), 3043–3051.
(13) Hou, X.; Li, S.; Frey, J.; Hong, X.; Ackermann, L. Machine Learning-Guided Yield Optimization for Palladaelectro-Catalyzed Annulation Reaction. *Chem* **2024**, *10* (7), 2283–2294.
(14) Xu, X. Y.; Liu, L. G.; Xu, L. C.; Zhang, S. Q.; Hong, X. Transfer Learning-Enabled Ligand Prediction for Ni-Catalyzed Atroposelective Suzuki-Miyaura Cross-Coupling Based on Mechanistic Similarity: Leveraging Pd Knowledge for Ni Discovery. *J. Am. Chem. Soc.* **2025**, *147* (18), 15318–15328.
(15) (a) Gao, W.; Raghavan, P.; Shprints, R.; Coley, C. W. Revealing the Relationship between Publication Bias and Chemical Reactivity with Contrastive Learning. *J. Am. Chem. Soc.* **2025**, *147* (10), 8959–8968. (b) Schleinitz, J.; Langevin, M.; Smail, Y.; Wehnert, B.; Grimaud, L.; Vuilleumier, R. Machine Learning Yield Prediction from NiCOlit, a Small-Size Literature Data Set of Nickel Catalyzed C–O Couplings. *J. Am. Chem. Soc.* **2022**, *144* (32), 14722–14730.
(16) (a) Torres, J. A. G.; Lau, S. H.; Anchuri, P.; Stevens, J. M.; Tabora, J. E.; Li, J.; Borovika, A.; Adams, R. P.; Doyle, A. G. A Multi-Objective Active Learning Platform and Web App for Reaction Optimization. *J. Am. Chem. Soc.* **2022**, *144* (43), 19999–20007. (b) Kelly, H. R.; Sreekumar, S.; Manee, V.; Cuomo, A. E.; Newhouse, T. R.; Batista, V. S.; Buono, F. Ligand-Based Principal Component Analysis Followed by Ridge Regression: Application to an Asymmetric Negishi Reaction. *ACS Catal.* **2024**, *14* (7), 5027–5038.
(17) Betinol, I. O.; Demchenko, A.; Reid, J. P. Evaluating Predictive Accuracy in Asymmetric Catalysis: A Machine Learning Perspective on Local Reaction Space. *ACS Catal.* **2025**, *15* (8), 6067–6077.
(18) (a) Osaro, E.; Karpinski, N.; Alornyo, S.; Ighalo, J. O. Artificial Intelligence in Materials Science and Chemistry: Past, Present and Future Trajectories. *Mater. Today Chem.* **2025**, *49*, No. 103115. (b) Wu, H.; Xie, S.; Li, X.; Zhao, Y.; Guo, L.; Tu, X.; Feng, F.; Wang, Q.; Lyu, J.; Feng, F.; Lu, C.; Fang, R.; Zhang, Q.; Li, X. Machine Learning for Catalysis: Bridging Data-Driven Discovery and Physical Insight. *Mater. Today Chem.* **2025**, *49*, No. 103051. (c) Vogiatzis, K. D.; Corminboeuf, C.; Nova, A.; Jorner, K.; Kästner, J.; Meuwly, M.; Schwaller, P.; Böttcher, V.; Drosou, M.; Fako, E.; Hoppe, H.; Ivkovic, Z.; Iwanojko, N.; Pantazis, D. A.; Schmid, S. P.; Szenes, K.; Tetenoire, A.; Reiher, M. Boosting Computational Catalysis and Chemical Reactivity with Artificial Intelligence. *J. Am. Chem. Soc.* **2026**, *148* (9), 9143–9155. (d) Schoepfer, A. A.; Laplaza, R.; Wodrich, M. D.; Waser, J.; Corminboeuf, C. Reaction-Agnostic Featurization of Bidentate Ligands for Bayesian Ridge Regression of Enantioselectivity. *ACS Catal.* **2024**, *14* (12), 9302–9312.
(19) (a) Cao, Z.; Falivene, L.; Poater, A.; Maity, B.; Zhang, Z.; Takasao, G.; Sayed, S. B.; Petta, A.; Talarico, G.; Oliva, R.; Cavallo, L. COBRA Web Application to Benchmark Linear Regression Models

Downloaded from pubs.acs.org/acscatcs/article-pdf/doi/10.1021/acscatal.6c05090/6849737/cs-2026-05090q.pdf by INSTITUTE OF ORGANIC CHEMISTRY PAS user on 26 September 2026

for Catalyst Optimization with Few-Entry Datasets. *Cell Rep. Phys. Sci.* **2025**, *6* (1), No. 102348. (b) Dalmau, D.; Sigman, M. S.; Alegre-Requena, J. V. Machine Learning Workflows beyond Linear Models in Low-Data Regimes. *Chem. Sci.* **2025**, *16* (19), 8555–8560. (c) Dalmau, D.; García-Abellán, S.; Alegre-Requena, J. V. Machine Learning in Homogeneous Catalysis: Basic Concepts and Best Practices. *ACS Catal.* **2025**, *16* (1), 1–11. (d) Schoepfer, A. A.; Laplaza, R.; Waser, J.; Corminboeuf, C. Feature Selection in Multivariate Linear Regression Models for Data-Efficient Reaction Optimization. *ChemRxiv* **2026**, *2026* (0601). (e) Dalmau, D.; Alegre-Requena, J. V. ROBERT: Bridging the Gap Between Machine Learning and Chemistry. *WIREs Comput. Mol. Sci.* **2024**, *14* (5), No. e1733. (f) Dalmau, D.; Jacot-Descombes, L.; Kalikadien, A.; Manzanilla, B.; Pidko, E. A.; Jorner, K.; Sigman, M. S.; Alegre-Requena, J. V. Cost-Effective Quantum-Mechanical Workflows for Molecular Machine Learning. *ACS Catal.* **2026**, *16*, 12565.

(20) (a) Samha, M. H.; Karas, L. J.; Vogt, D. B.; Odogwu, E. C.; Elward, J.; Crawford, J. M.; Steves, J. E.; Sigman, M. S. Predicting Success in Cu-Catalyzed C–N Coupling Reactions Using Data Science. *Sci. Adv.* **2026**, *10* (3), No. eadn3478. (b) Haas, B. C.; Kalyani, D.; Sigman, M. S. Applying Statistical Modeling Strategies to Sparse Datasets in Synthetic Chemistry. *Sci. Adv.* **2026**, *11* (1), No. eadt3013.

(21) Kozlowski, M. C. On the Topic of Substrate Scope. *Org. Lett.* **2022**, *24* (40), 7247–7249.

(22) Jaszczewska-Adamczak, J. A.; Mlynarski, J. Asymmetric Epoxidation of Enones Promoted by Dinuclear Magnesium Catalyst. *Adv. Synth. Catal.* **2021**, *363* (17), 4247–4255.

(23) Jaszczewska-Adamczak, J. A.; Baczewska, P.; Bujok, R.; Mlynarski, J. Magnesium-Catalyzed Asymmetric Thia-Michael Addition to α,β-Unsaturated Ketones. *Adv. Synth. Catal.* **2024**, *366* (6), 1412–1421.

(24) (a) Pietrzak, T.; Kubisiak, M.; Justyniak, I.; Zelga, K.; Bojarski, E.; Tratkiewicz, E.; Ochal, Z.; Lewiński, J. Oxygenation Chemistry of Magnesium Alkyls Incorporating β-Diketiminate Ligands Revisited. *Chem. Eur. J.* **2016**, *22* (49), 17776–17783. (b) Lewiński, J.; Ochal, Z.; Bojarski, E.; Tratkiewicz, E.; Justyniak, I.; Lipkowski, J. First Structurally Authenticated Zinc Alkylperoxide: A Model System for the Epoxidation of Enones. *Angew. Chem., Int. Ed.* **2003**, *42* (38), 4643–4646. (c) Raheem Keeri, A.; Justyniak, I.; Jurczak, J.; Lewiński, J. Quest for Efficient Catalysts Based on Zinc Tert-Butyl Peroxides for Asymmetric Epoxidation of Enones: C_2- vs C_1-Symmetric Auxiliaries. *Adv. Synth. Catal.* **2016**, *358* (6), 864–868.

(25) Zheng, K.; Liu, X.; Feng, X. Recent Advances in Metal-Catalyzed Asymmetric 1,4-Conjugate Addition (ACA) of Nonorganometallic Nucleophiles. *Chem. Rev.* **2018**, *118* (16), 7586–7656.

(26) Moriwaki, H.; Tian, Y. S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *J. Cheminform.* **2018**, *10* (1), 4.

(27) (a) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular Fingerprint Similarity Search in Virtual Screening. *Methods* **2015**, *71* (C), 58–63. (b) Landrum, G.; et al. RDKit: Open-Source Cheminformatics Software, version 2025.09.3. *Zenodo*; **2025**.

(28) Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *arXiv* **2020**, No. arXiv:2010.09885.

(29) Byadi, S.; Gantzer, P.; Gimadiev, T.; Sidorov, P. DOPtools: A Python Platform for Descriptor Calculation and Model Optimization. *Digit. Discov.* **2025**, *4* (5), 1188–1198.

(30) (a) Jorner, K.; Turcani, L. Kjelljorner/Morfeus: V0.7.2. *Zenodo*; **2022**. (b) Verloop, A.; Hoogenstraaten, W.; Tipker, J. Chapter 4. Development and Application of New Steric Substituent Parameters in Drug Design. In *Medicinal Chemistry: A Series of Monographs*; Ariëns, E. J., Eds.; Academic Press: Amsterdam, **1976**, *11*, pp 165–207. (c) Harper, K. C.; Bess, E. N.; Sigman, M. S. Multidimensional Steric Parameters in the Analysis of Asymmetric Catalytic Reactions. *Nat. Chem.* **2012**, *4* (5), 366–374.

(31) Gasteiger, J.; Marsili, M. A New Model for Calculating Atomic Charges in Molecules. *Tetrahedron Lett.* **1978**, *19* (34), 3181–3184.

(32) Nisbet, R.; Elder, J.; Miner, G. *Handbook of Statistical Analysis and Data Mining Applications*; Academic Press, **2009**.

(33) Ho, T. K. Random Decision Forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*; **1995**, *1*, pp 278–282.

(34) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, **2016**, pp 785–794.

(35) Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29* (5), 1189–1232.

(36) Cox, D. R. The Regression Analysis of Binary Sequences. *J. R. Stat. Soc. B* **1958**, *20* (2), 215–232.

(37) Hoerl, A. E.; Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **1970**, *12* (1), 55–67.

(38) von Winterfeldt, D.; Edwards, W. *Decision Analysis and Behavioral Research*; Cambridge University Press, **1986**.

(39) Fix, E.; Hodges, J. L. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. *Int. Stat. Rev.* **1989**, *57* (3), 238–247.

(40) Awad, M.; Khanna, R. Efficient Learning Machines. In *Support Vector Regression*; Apress: Berkeley, **2015**, pp 67–80.

(41) Gao, B.; Chang, Y.; Tang, W. Prediction of the Enantiomeric Excess Value for Asymmetric Transfer Hydrogenation Based on Machine Learning. *Org. Chem. Front.* **2023**, *10* (6), 1456–1462.

(42) Rani, N.; Kumar, R.; Mazumder, S. AI-Driven Discovery of Asymmetric Pauson–Khand Reactions: A New Toolbox in a Synthetic Chemist's Treasure. *J. Phys. Chem. A* **2024**, *128* (48), 10452–10463.

(43) Hoque, A.; Jain, N.; Chenna, D.; Sunoj, R. B. Machine Learning Models for Catalytic Asymmetric Reactions of Simple Alkenes: From Enantioselectivity Predictions to Chemical Insights. *Digit. Discov.* **2026**, *5* (4), 1718–1735.

(44) (a) Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*; Curran Associates, Inc.: Red Hook, NY, **2017**, pp 4765–4774. (b) Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2020**, *2* (1), 56–67.

Downloaded from pubs.acs.org/acscatcs/article-pdf/doi/10.1021/acscatal.6c05090/6849375/cs-2026-05090q.pdf by INSTITUTE OF ORGANIC CHEMISTRY PAS user on 26 September 2026

10. Oświadczenia Autorów

Instytut Chemii Organicznej
Polskiej Akademii Nauk

mgr Paulina Baczewska
Zespół XIII IChO PAN
paulina.baczewska@icho.edu.pl

Oświadczam, że mój wkład w powstanie poniższych publikacji:

- **P. Baczewska**, M. Kulczykowski, B. Zambroń, J. Jaszczewska-Adamczak, Z. Pakulski, R. Roszak, B. A. Grzybowski, J. Mlynarski
  Angew. Chem. Int. Ed. **2024**, 63, e202318487
  Machine Learning Algorithm Guides Catalyst Choices for Magnesium-Catalyzed Asymmetric Reactions
  DOI: 10.1002/anie.202318487

polegał na wiodącym wkładzie oraz nadzorowaniu ręcznego ekstrahowania danych z publikacji, opracowaniu danych i ogólnym tworzeniu bazy danych, współudziale w rozwoju modelu uczenia maszynowego i aplikacji sieciowej, przeprowadzeniu wszystkich eksperymentów walidacyjnych opisanych w artykule oraz współudziale w napisaniu manuskryptu publikacji, napisaniu materiałów uzupełniających (SI) oraz intensywna praca nad poprawkami oraz listem do redakcji.

- **P. Baczewska**, D. Nowak, J. Jaszczewska-Adamczak, R. A. Bachorz, M. Hoffmann, J. Mlynarski
  ACS Catal. **2026**,
  Predicting Reactivity and Enantioselectivity from Small Experimental Datasets: A Case Study in Asymmetric Magnesium Catalysis
  DOI: [nadany po publikacji]

polegał na zebraniu i opracowaniu zbiorów danych do budowy modeli, współudziale w budowaniu oryginalnej metody (ocena jakości modeli dla katalizy, zaproponowanie zastosowania reprezentacji chemicznych, które pojawiły się w finalnych modelach – deskryptory Circus i ChyLine), interpretacji wyników modelowania i analizie zakresu stosowalności modeli, zaplanowaniu i przeprowadzeniu większości weryfikacji eksperymentalnej przewidywań modeli (współpraca przy epoksydacji chalkonów i całość addycji tia-Michaela), współudziale w pisaniu manuskryptu publikacji i materiałów uzupełniających (SI) oraz intensywna praca nad poprawkami po uwagach od recenzentów.

Warszawa, dnia 18.01.2026 .....
(podpis)

Prof. dr hab. Jacek Młynarski

Oświadczam, że mój wkład w powstanie powyższych publikacji polegał na współtworzeniu koncepcji badań, pozyskaniu finansowania (grant OPUS NCN 2021/41/B/ST4/01617), kierowaniu projektem badawczym, współpracy nad pisaniem manuskryptów publikacji, kontakcie z wydawcą oraz pracą nad listem do redakcji.

Warszawa, dnia 15/09/2025 .....
(podpis)

OŚWIADCZENIE

Oświadczamy, że w artykule pt. „Machine Learning Algorithm Guides Catalyst Choices for Magnesium-Catalyzed Asymmetric Reactions", opublikowanym w czasopiśmie Angewandte Chemie International Edition (2024, 63, e202318487, DOI: 10.1002/anie.202318487), wkład wymienionych Autorów w przygotowanie publikacji jest następujący:

LP	Imię i nazwisko Autora	Wkład merytoryczny Autora	Podpis Autora
1	dr Michał Kulczykowski	Współpraca nad opracowaniem danych i tworzeniu bazy danych, wiodący udział w opracowaniu opublikowanego modelu i aplikacji sieciowej oraz współpraca nad pisaniem manuskryptu i opracowaniu poprawek od redakcji.	
2	dr hab. Bartosz Zambroń	Współudział w procesie ręcznego zbierania danych z publikacji i tworzeniu baz danych, synteza ligandów Fenga stosowanych w eksperymentach walidacyjnych	
3	dr inż. Joanna Jaszczewska-Adamczak	Współudział w procesie ręcznego zbierania danych z publikacji i tworzeniu baz danych	
4	Prof. Zbigniew Pakulski	Synteza ligandów CPA stosowanych w eksperymentach walidacyjnych	
5	dr Rafał Roszak	Współpraca nad tworzeniem modelu na wczesnym etapie koncepcji oraz współpraca nad poprawkami do redakcji	
6	prof. dr hab. Bartosz A. Grzybowski	Współtworzenie koncepcji badań i nadzorowanie projektu, współudział w pisaniu manuskryptu publikacji	
7	prof. dr hab. Jacek Młynarski	Współtworzenie koncepcji badań i nadzorowanie projektu, współudział w pisaniu manuskryptu publikacji	

OŚWIADCZENIE

Oświadczamy, że w artykule pt. „Predicting Reactivity and Enantioselectivity from Small Experimental Datasets: A Case Study in Asymmetric Magnesium Catalysis”, zaakceptowanym do publikacji w czasopiśmie ACS Catalysis, wkład wymienionych Autorów w przygotowanie publikacji jest następujący:

LP	Imię i nazwisko Autora	Wkład merytoryczny Autora	Podpis Autora
1	mgr Damian Nowak	Zebranie, kuracja i przetwarzanie danych do budowy modeli, opracowanie kodu i realizacja analiz opartych na uczeniu maszynowym, dobór, optymalizacja, walidacja i ocena modeli predykcyjnych, interpretacja wyników i przygotowanie wizualizacji oraz współudział w przygotowaniu manuskryptu, materiałów uzupełniających i poprawek po recenzjach.	Podpisano przez/ Signed by: DAMIAN ALEKSANDER NOWAK Data/ Date: 14.09.2026 11:09 mSzafir
2	dr inż. Joanna Jaszczewska-Adamczak	Współudział w prowadzeniu walidacji eksperymentalnej w zakresie epoksydacji i syntez substratów, współudział w przygotowaniu materiałów uzupełniających.	PODPIS ZAUFANY JOANNA ADRIANA JASZCZEWSKA-ADAMCZAK 15.09.2026 14:30:16 GMT+0200 Dokument podpisany elektronicznie podpisem zaufanym
3	dr Rafał Adam Bachorz	Dyskusja wyników, współudział przy opracowywaniu oryginalnej metody oraz manuskryptu, współpraca nad poprawkami manuskryptu.	PODPIS ZAUFANY RAFAŁ BACHORZ 14.09.2026 16:47:38 GMT+0200 Dokument podpisany elektronicznie podpisem zaufanym
4	prof. dr hab. Marcin Hoffmann	Dyskusja wyników, współudział przy opracowywaniu oryginalnej metody oraz manuskryptu.	PODPIS ZAUFANY MARCIN MACIEJ HOFFMANN 14.09.2026 18:07:57 GMT+0200 Dokument podpisany elektronicznie podpisem zaufanym
5	prof. dr hab. Jacek Młynarski	Dyskusja wyników, współudział przy opracowywaniu oryginalnej metody oraz manuskryptu, opracowanie poprawek manuskryptu i listu do redakcji.	Jacek Jan Młynarski Elektronicznie podpisany przez Jacek Jan Młynarski Data: 2026.09.15 15:20:50 +02'00'