Open Applied Science

https://openappscience.science

Data-Driven Screening of Chromium Catalysts for Selective Ethylene Oligomerization

Margaux Vidal

Department of Chemistry, College of Science and Technology, Université de Bordeaux, Bordeaux, Nouvelle-Aquitaine, France

Abstract

The selective oligomerization of ethylene to produce linear alpha olefins, specifically 1-hexene and 1-octene, represents one of the most vital homogeneous catalytic processes in the modern petrochemical industry. While chromium-based complexes coordinated by heteroatomic bidentate and tridentate ligands dominate industrial applications, the discovery of novel catalyst architectures with superior activity, thermal stability, and product selectivity remains a formidable challenge. Traditional catalyst development relies heavily on empirical trial-and-error synthesis coupled with computationally expensive quantum mechanical modeling of complex metallacycle pathways. This paper presents a comprehensive data-driven screening framework that integrates organometallic domain knowledge, density functional theory descriptor extraction, and machine learning algorithms to accelerate the discovery of high-performance chromium catalysts. By curating an extensive dataset of structurally diverse bis(phosphino)amine and mixed-donor chromium complexes, a robust set of steric, electronic, and topological descriptors was engineered to capture the subtle ligand effects governing the competition between metallacycloheptane and metallacyclononane intermediates. Multiple supervised learning models were evaluated, with gradient boosted decision trees demonstrating exceptional predictive accuracy for both turnover frequency and C6 versus C8 selectivity. Subsequent high-throughput virtual screening of a combinatorially generated library of twelve thousand candidate ligands, guided by multi-objective Pareto optimization, identified several previously unexplored asymmetric ligand scaffolds predicted to exhibit enhanced 1-octene selectivity and suppressed polyethylene wax formation. Interpretability analysis utilizing Shapley additive explanations further elucidated the non-linear interplay between phosphorus substituent buried volume and metal center natural population charge, establishing a quantitative structure-selectivity relationship for rational catalyst design.

Keywords: ethylene oligomerization, chromium catalysis, machine learning, quantitative structure-property relationships

1. Introduction

1.1 Industrial Significance of Linear Alpha Olefins

Linear alpha olefins constitute a cornerstone class of commodity chemicals with vast industrial utility across the global chemical manufacturing sector [1]. Among the homologous series of linear alpha olefins, the shorter-chain oligomers, specifically 1-butene, 1-hexene, and 1-octene, hold paramount economic value due to their extensive consumption as comonomers in the production of linear low-density polyethylene and high-density polyethylene. The incorporation of 1-hexene and 1-octene into the polyethylene backbone introduces short-chain branching, which disrupts crystallinity and substantially improves the mechanical properties of the resulting polymer, including tear resistance, tensile strength, impact durability, and environmental stress cracking resistance. Historically, the industrial production of linear alpha olefins relied predominantly on unselective oligomerization routes, such as the Shell Higher Olefin Process and various aluminum-alkyl catalyzed chain growth processes. These conventional technologies inevitably generate a broad Schulz-Flory or Poisson distribution of even-numbered carbon olefins ranging from C4 to C30+. While these full-range processes remain operationally viable when market demand aligns with the statistical

distribution of products, the rapid expansion of high-performance polyethylene grades has created an acute market imbalance characterized by a surging, disproportionate demand for pure 1-hexene and 1-octene fractions.

Consequently, the petrochemical industry has shifted its strategic focus toward on-purpose, selective ethylene oligomerization technologies capable of bypassing statistical product distributions to yield exclusively trimers or tetramers of ethylene. Selective ethylene trimerization to 1-hexene and tetramerization to 1-octene offer profound economic and environmental advantages over traditional full-range processes. By targeting specific carbon numbers with high selectivity, chemical producers can eliminate the capital-intensive and energy-demanding fractional distillation trains otherwise required to separate complex mixtures of alpha olefins, internal olefins, and branched isomers. Furthermore, selective catalytic systems minimize the co-production of lower-value heavy olefin fractions, thereby maximizing carbon efficiency and feedstock utilization. Despite these advantages, achieving simultaneous control over catalytic activity, product selectivity toward 1-octene over 1-hexene, and suppression of undesirable polymeric byproducts remains an intricate chemical problem. Particularly troublesome in industrial loop and continuous stirred-tank reactors is the formation of insoluble high-molecular-weight polyethylene wax, which causes severe reactor fouling, heat transfer deterioration, and premature unit shutdowns. A contrasting application concerns olive-leaf phenol extraction and oil enrichment, with distinct physical and validation requirements [2].

1.2 Evolution of Chromium-Based Catalytic Systems

Among the transition metals investigated for selective ethylene oligomerization, chromium occupies a unique and unrivaled position [3] [4]. Whereas nickel, titanium, and zirconium complexes typically operate via a Cossee-Arlman migratory insertion mechanism that yields a statistical Schulz-Flory distribution of oligomers, chromium catalysts facilitate a distinct oxidative coupling pathway involving cyclic organometallic intermediates that naturally favor selective trimerization and tetramerization. The landmark discovery of the Phillips pyrrolide-based chromium system for ethylene trimerization demonstrated the commercial viability of on-purpose 1-hexene manufacturing. Subsequent academic and industrial breakthroughs led to the development of chelating heteroatomic ligands, most notably the bis(phosphino)amine, or PNP, ligand scaffold introduced by researchers at Sasol, which revolutionized the field by enabling the first highly selective catalytic tetramerization of ethylene to 1-octene.

Following the discovery of the prototypical N,N-bis(diphenylphosphino)alkylamine chromium systems, an expansive array of ligand architectures has been synthesized and evaluated [5]. These include diphosphinoamine derivatives with varied nitrogen and phosphorus substituents, bis(sulfanylmethyl)pyridine and related SNS pincer ligands, phosphino-oxazoline motifs, and carbon-bridged PCCP and PCN diphosphine frameworks. Small structural perturbations within these ligands frequently induce dramatic shifts in catalytic behavior. For instance, introducing ortho-methoxy or ortho-fluoro groups onto the aryl rings of a PNP ligand can switch the dominant reaction pathway entirely from tetramerization to trimerization, or conversely, lead to rapid catalyst deactivation. Furthermore, the performance of these precatalysts is heavily influenced by the nature of the cocatalyst, typically modified methylaluminoxane, and the specific reaction environment, including solvent polarity, ethylene partial pressure, and reaction temperature. Because the active chromium species exist as highly air-sensitive, paramagnetic complexes in multiple accessible oxidation states, in situ spectroscopic characterization of the true catalytic resting states and transition states remains exceedingly difficult.

1.3 Emergence of High-Throughput Virtual Screening

Given the vast combinatorial space of potential ancillary ligands and the sensitivity of chromium catalysts to subtle stereoelectronic variations, traditional empirical discovery methods face severe scalability bottlenecks. Conventional laboratory workflows require the multi-step synthesis of air-sensitive phosphine or sulfur ligands, complexation with chromium precursors, and high-pressure autoclave testing with pyrophoric aluminum alkyl activators. While high-throughput parallel reactor systems have accelerated experimental screening to some degree, physical experimentation remains constrained by material costs, synthesis time, and human labor [6] [7]. Over the past decade, quantum mechanical modeling utilizing

density functional theory has provided invaluable mechanistic insights into the energetics of ethylene coordination, oxidative coupling, and beta-hydride transfer steps. However, full catalytic cycle calculations for open-shell transition metal complexes across multiple spin surfaces are computationally demanding, requiring hundreds of processor hours per single ligand candidate and precluding the brute-force quantum chemical screening of thousands of hypothetical structures. Sequential feature-selection algorithms offer a methodological basis for comparing candidate descriptor subsets [8].

To bridge the gap between slow, accurate quantum mechanical calculations and rapid, large-scale chemical space exploration, data-driven methodologies and machine learning have emerged as transformative paradigms in homogeneous catalysis [9] [10]. By translating molecular structures into quantitative steric, electronic, and topological descriptors, supervised machine learning algorithms can learn complex, non-linear mappings between catalyst features and experimental or computed performance metrics. Once trained and validated on representative datasets, these surrogate models can evaluate tens of thousands of virtual catalyst candidates in mere seconds, identifying optimal regions of the chemical space that warrant targeted experimental synthesis. In this study, we establish a comprehensive, end-to-end data-driven screening pipeline specifically tailored for chromium-catalyzed selective ethylene oligomerization. By harmonizing historical experimental data with targeted density functional theory descriptor generation, we construct predictive machine learning models capable of simultaneously forecasting catalytic turnover frequency and C8 versus C6 product selectivity, ultimately uncovering novel ligand designs that push the boundaries of selective olefin synthesis. LASSO regression provides a sparsity-based option for selecting predictive descriptors [11].

2. Mechanistic and Structural Background

2.1 Metallacycle Intermediates and Catalytic Cycles

The rational design of descriptors for machine learning models in catalysis necessitates a rigorous understanding of the underlying reaction mechanism. Unlike conventional olefin polymerization catalysts that propagate via linear alkyl chains, chromium-catalyzed selective ethylene oligomerization proceeds through a ring-expansion metallacycle mechanism first proposed by Briggs and later substantiated through isotopic labeling experiments and computational studies. Upon activation of a chromium(III) precatalyst with an excess of methylaluminoxane in an inert hydrocarbon solvent, abstraction of anionic halide ligands and reduction processes generate a coordinatively unsaturated, cationic low-valent chromium active species, generally accepted to reside in the formal chromium(I) or chromium(II) oxidation state depending on the specific ligand architecture. The catalytic cycle initiates with the coordination of two ethylene molecules to this low-valent metal center, followed by an oxidative coupling step that forms a five-membered chromacyclopentane intermediate, wherein the formal oxidation state of the metal increases by two units. A contrasting application concerns near-infrared monitoring of botanical processing, with distinct physical and validation requirements [12].

Crucially, the geometric constraints of the five-membered metallacyclopentane ring prevent facile beta-hydride elimination, thereby suppressing the premature liberation of 1-butene. Instead, the metallacyclopentane intermediate preferentially coordinates a third molecule of ethylene, which inserts into one of the chromium-carbon sigma bonds to yield a seven-membered chromacycloheptane intermediate. At this juncture in the catalytic cycle, a critical mechanistic bifurcation occurs that dictates the ultimate selectivity between 1-hexene and 1-octene. Along the trimerization pathway, the flexible seven-membered ring undergoes an intramolecular beta-hydride transfer to the opposite alpha-carbon, either directly through a concerted transition state or via a chromium-hydride-alkenyl intermediate, followed by reductive elimination to release 1-hexene and regenerate the active low-valent chromium species. Conversely, along the tetramerization pathway, the chromacycloheptane intermediate coordinates a fourth molecule of ethylene and undergoes an additional ring expansion via migratory insertion to form a nine-membered chromacyclononane species. Subsequent beta-hydride transfer and reductive elimination from this nine-membered ring liberate 1-octene and close the catalytic cycle. The complete mechanistic network, alongside the extraction of key stereoelectronic parameters that govern these competing pathways, is illustrated in

Figure 1. The application of chromatographic measurement of plant glycosides requires its own assumptions when comparing modeling approaches [13].

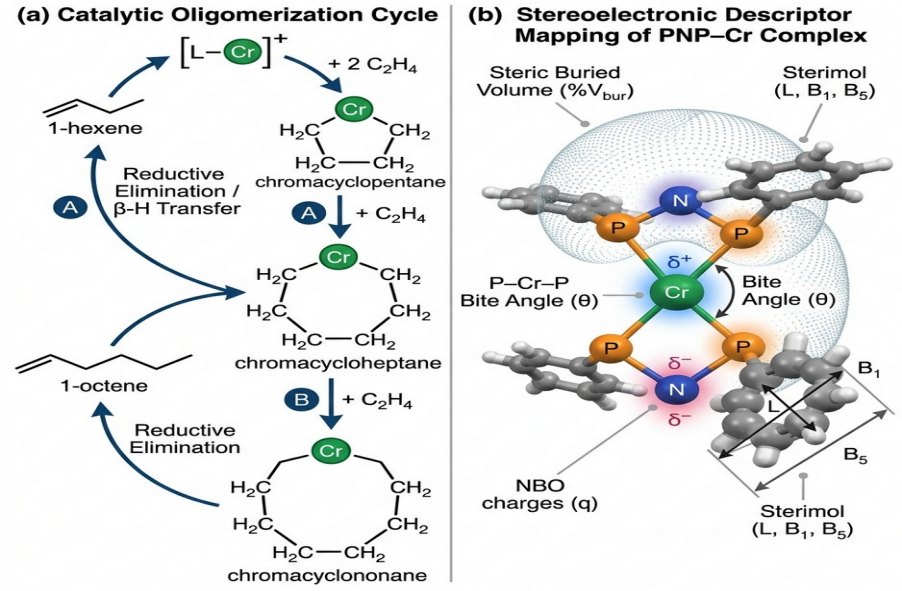


Figure 1: Mechanistic Bifurcation of Chromium-Catalyzed Ethylene Oligomerization and Stereoelectronic Descriptor Mapping

The relative rates of beta-hydride transfer versus ethylene insertion at the chromacycloheptane stage are governed by the free energy difference between their respective transition states [14]. Furthermore, undesired side reactions can branch off from this primary manifold. For instance, if the nine-membered ring continues to insert ethylene molecules faster than it undergoes elimination, high-molecular-weight polyethylene chains are formed. Alternatively, isomerization processes or re-insertion of already formed 1-hexene into the growing metallacycle results in branched decenes and tetradecenes, as well as internal olefins and cyclopentane derivatives, all of which degrade product purity.

2.2 Ligand Architecture and Electronic Modulation

The competition between ring expansion and ring elimination at the seven-membered metallacycle stage is exquisitely sensitive to the steric and electronic environment imposed by the ancillary ligand [15]. In the widely studied N,N-bis(phosphino)amine chromium systems, the P-N-P backbone forms a strained four-membered chelate ring upon bidentate coordination to the chromium center. This small bite angle, typically ranging between 65 and 73 degrees, leaves a substantial open coordination wedge at the metal center, which facilitates the simultaneous coordination of ethylene monomer alongside the growing metallacycle. When bulkier substituents, such as isopropyl, cyclohexyl, or tert-butyl groups, are attached to the central nitrogen atom, steric repulsion pushes the flanking diarylphosphino groups toward the chromium center. This steric compression restricts the conformational flexibility of the seven-membered chromacycloheptane ring, destabilizing the agostic C-H interaction required for beta-hydride elimination and thereby favoring the bimolecular insertion of a fourth ethylene molecule to yield 1-octene.

However, steric bulk is a double-edged sword. Excessive steric crowding around the axial coordination sites of the chromium center hinders the approach of incoming ethylene monomers and destabilizes the active catalyst, leading to reduced turnover frequencies and accelerated decomposition via binuclear deactivation pathways [16]. Electronic effects play an equally decisive role. Electron-donating substituents on the phosphorus atoms increase the electron density at the chromium center, which enhances back-donation into the pi-antibonding orbitals of coordinated ethylene and lowers the activation barrier for oxidative coupling. Conversely, ortho-substituted aryl rings bearing pendant donor groups, such as ortho-methoxyphenyl moieties, can engage in hemilabile coordination with the chromium center. In these hemilabile systems, the oxygen atom transiently coordinates to the metal during the catalytic cycle, blocking the fourth coordination site needed for ethylene uptake by the chromacycloheptane ring and enforcing a

strict trimerization pathway that yields 1-hexene with selectivities exceeding ninety-eight percent. Beyond PNP ligands, extending the backbone bridge length to create P-C-C-P or P-N-P-O frameworks alters the bite angle and chelate ring flexibility, providing additional handles to fine-tune the balance between catalytic activity, thermal robustness, and C8 versus C6 product ratios.

2.3 Limitations of Conventional Empirical Discovery

Despite the wealth of qualitative mechanistic insights accumulated over two decades of organometallic research, translating these principles into quantitative design rules for new catalysts remains exceptionally difficult. Empirical structure-activity relationships derived from small, localized experimental studies frequently fail when extrapolated to broader ligand classes. This breakdown occurs because steric and electronic effects in transition metal complexes are inherently entangled; modifying a functional group on a phosphorus or nitrogen atom simultaneously alters the inductive electronic donation, the conformational preferences of the ligand backbone, the dispersion interactions stabilizing the transition states, and the steric accessibility of the metal center. A contrasting application concerns excitonic interactions in suspended semiconductors, with distinct physical and validation requirements [17]. Random forests provide a tree-ensemble modeling option for descriptor-response relationships [18].

Furthermore, historical catalyst testing data scattered across academic literature and patent databases suffer from significant variance in experimental testing protocols [19]. Variations in reactor geometry, mass transfer limitations of ethylene gas into liquid solvent phases, trace impurities in methylaluminoxane batches, and differences in catalyst activation temperatures introduce systematic noise that obscures subtle ligand effects. Consequently, human researchers struggle to discern multi-dimensional correlations across hundreds of disparate catalyst structures. A data-driven approach that standardizes experimental metrics, represents catalysts through physically meaningful quantum chemical and geometric descriptors, and applies non-linear statistical learning algorithms offers a rigorous pathway to overcome these human cognitive and empirical limitations [20].

3. Data Curation and Feature Engineering Methodology

3.1 Compilation of Experimental and Computational Datasets

The foundation of any robust machine learning framework in chemical discovery is a high-quality, standardized dataset that spans a representative domain of the chemical space [21]. For this investigation, a comprehensive literature and patent mining campaign was conducted to compile experimental catalytic performance data for homogeneous chromium oligomerization catalysts reported over the past twenty years. To minimize confounding variance arising from disparate reactor engineering conditions, strict inclusion criteria were enforced during data curation. Only catalytic runs utilizing modified methylaluminoxane or trimethylaluminum-depleted methylaluminoxane as the activator, operated in non-polar hydrocarbon solvents such as cyclohexane, methylcyclohexane, or heptane, within a temperature window of 40 to 80 degrees Celsius and an ethylene pressure range of 30 to 50 bar, were retained for the primary training database.

The curated dataset comprises 480 unique chromium complexes spanning six primary ligand families: canonical N-bridged PNP ligands, C-bridged PCCP ligands, N,N-bridged PNPN ligands, phosphino-oxazoline ligands, SNS tridentate pincer ligands, and hemilabile PNPO scaffolds. For each data entry, the experimental catalytic activity, standardized to units of kilograms of oligomer produced per gram of chromium per hour, and the mass percentage distributions of 1-hexene, 1-octene, internal hexenes, C10+ heavy oligomers, and polyethylene wax were recorded. Because reaction temperature, ethylene pressure, aluminum-to-chromium molar ratio, and reaction duration still varied within the accepted operational windows, these four continuous experimental process parameters were explicitly retained as contextual input features alongside the molecular descriptors of the catalysts. This strategy enables the machine learning algorithms to decouple intrinsic ligand electronic and steric effects from extrinsic reaction engineering variables.

3.2 Extraction of Quantum Mechanical and Steric Descriptors

To represent the 480 curated chromium complexes in a format amenable to machine learning while preserving deep chemical fidelity, a standardized computational chemistry protocol was executed to extract three-dimensional geometric, steric, and quantum mechanical descriptors. Rather than computing the entire multi-step catalytic cycle for every ligand, which would defeat the purpose of rapid screening, descriptor extraction was performed on a uniform reference state: the cationic, mononuclear chromium(III) dichloride mono-solvent complex, alongside the uncoordinated free ligand geometry. Geometry optimizations were carried out using density functional theory with an unrestricted dispersion-corrected hybrid functional and a def2-SVP basis set for main group elements coupled with a def2-TZVP basis set for the chromium center, assuming a high-spin quartet ground state for the d3 chromium(III) configuration.

From the converged equilibrium geometries, a diverse pool of 64 molecular descriptors was systematically calculated for each complex. Electronic descriptors were derived from Natural Bond Orbital analysis and frontier molecular orbital theory, capturing the natural population charge on the chromium center, the average natural charge on the coordinating donor atoms, the energy of the highest occupied and lowest unoccupied alpha and beta spin orbitals, the global electrophilicity index, and the metal-ligand Wiberg bond indices [22]. Geometric and steric descriptors were computed using automated structural analysis routines to quantify the local environment around the catalytic pocket. These included the donor-chromium-donor bite angle, average chromium-donor bond lengths, the percentage buried volume within spheres of radii ranging from 3.0 to 5.0 Angstroms centered on the chromium atom, quadrant-resolved buried volume differentials to measure steric asymmetry, and multidimensional Sterimol parameters describing the length, minimum width, and maximum width of the substituents attached to the nitrogen and phosphorus donor atoms. In addition, two-dimensional topological fingerprints and autocorrelation vectors weighted by Pauling electronegativity and atomic polarizability were generated directly from the molecular connection tables to facilitate rapid pre-screening of virtual libraries prior to three-dimensional coordinate generation. A structured summary of the primary descriptor categories utilized in the modeling pipeline is presented in Table 1.

Table 1: Summary of curated molecular, steric, and electronic descriptors utilized for model training

Descriptor Category	Specific Feature Examples	Physical and Mechanistic Relevance	Computational Source
Electronic Properties	Cr natural charge, donor atom charges, HOMO and LUMO energies, Wiberg bond indices	Controls ethylene pi-backdonation, oxidative coupling barriers, and metal oxidation state stability	DFT Natural Bond Orbital and frontier orbital analysis
Global Steric Bulk	Percentage buried volume at 3.5 A, 4.5 A, andtotal catalyst molecular volume	Governs monomer accessibility to the active site and suppresses binuclear deactivation pathways	SambVca geometric sphere integration on DFT structures
Local Asymmetric Sterics	Quadrant buried volume variance, Sterimol B1, B5, and L of N and P substituents	Dictates ring conformation of chromacycloheptane and competition between beta-H transfer and insertion	3D coordinate vector projection and Sterimol analysis
Chelate Geometry	P-Cr-P bite angle, N-P-Cr bond angles, Cr-P and Cr-N equilibrium bond lengths	Modulates the open coordination wedge angle and strain energy of the metallacycle intermediates	Converged DFT equilibrium Cartesian coordinates
Topological Indices	Morgan circular fingerprints, Moreau-Broto autocorrelations, Kier flexibility index	Captures broad structural motifs and enables ultra-fast pre-filtering of virtual combinatorial libraries	2D molecular graph connection table calculations
Process Conditions	Reaction temperature, ethylene pressure, Al to Cr molar ratio, run duration	Accounts for monomer solubility, thermal catalyst decay, and cocatalyst activation efficiency	Curated experimental literature and patent records

3.3 Feature Selection and Dimensionality Reduction Protocols

Applying machine learning algorithms directly to a high-dimensional feature space containing 68 combined molecular and process variables on a dataset of 480 observations introduces a significant risk of overfitting

and multicollinearity. Many quantum chemical and steric descriptors exhibit strong internal correlations; for instance, the percentage buried volume at a radius of 3.5 Angstroms correlates heavily with the buried volume at 4.0 Angstroms, and the Wiberg bond indices correlate inversely with equilibrium metal-ligand bond lengths. To distill the descriptor pool into a compact, orthogonal, and physically interpretable subset, a rigorous multi-stage feature selection protocol was implemented. The application of data representation and grouping requires its own assumptions when comparing modeling approaches [23].

In the first stage, features exhibiting near-zero variance across the dataset were removed, followed by a pairwise Pearson correlation filtration step [24]. Whenever two descriptors displayed an absolute correlation coefficient exceeding 0.85, the descriptor with the higher average correlation against all other remaining features in the matrix was systematically eliminated. This pruning reduced the initial feature pool from 68 variables down to 34 relatively independent descriptors. In the second stage, recursive feature elimination coupled with cross-validated random forest regression was performed independently for the two primary target variables: log-transformed catalytic activity and the C8-to-C6 mass selectivity ratio. By iteratively training models, ranking features by their permutation importance scores, and discarding the least informative variables, the optimal feature subset size was identified at the inflection point of the cross-validation error curve. This procedure yielded a final refined set of 16 core descriptors that captured over ninety-four percent of the predictive variance while ensuring high computational efficiency and model interpretability.

4. Machine Learning Architecture and Screening Pipeline

4.1 Predictive Algorithm Selection and Hyperparameter Tuning

Because the functional relationship between catalyst stereoelectronic descriptors and oligomerization performance involves highly non-linear threshold effects, such as sudden selectivity switches induced by hemilabile donor coordination or steric congestion limits, a diverse ensemble of linear and non-linear supervised regression algorithms was benchmarked. The evaluated model architectures encompassed Ridge Regression and Elastic Net as regularized linear baselines, Support Vector Regression utilizing radial basis function kernels, Gaussian Process Regression with Matern covariance kernels, Random Forest Regression, and Extreme Gradient Boosting decision trees. Separate predictive models were trained for three critical target outputs: the logarithm of total catalytic turnover frequency, the selectivity ratio of 1-octene to 1-hexene, and the weight percentage of undesired polyethylene byproduct. Deep Q-EP was developed for data with uneven local regularity [25]. Such regularity is a relevant consideration when comparing surrogate families for the threshold-like selectivity changes described here; catalyst-specific accuracy must be established from the chemical observations themselves.

To prevent data leakage and assess the true generalization capacity of the models on unseen ligand scaffolds, both standard ten-fold cross-validation and scaffold-split validation were conducted [26]. In the scaffold-split regime, complexes sharing an identical core ligand backbone were grouped into the same validation fold, forcing the algorithms to extrapolate across structural motifs rather than merely interpolating within a homologous series of para-substituted aryl rings. Hyperparameter optimization for each algorithm was executed via Bayesian optimization using a tree-structured Parzen estimator over five hundred iterations, minimizing the root mean squared error across the inner cross-validation loops. All continuous input descriptors were standardized to zero mean and unit variance using scaling parameters derived strictly from the training partitions. Figure 2 illustrates the end-to-end architecture of the data-driven workflow, detailing the progression from dataset curation and descriptor calculation through machine learning model training to high-throughput virtual screening and multi-objective Pareto front selection.

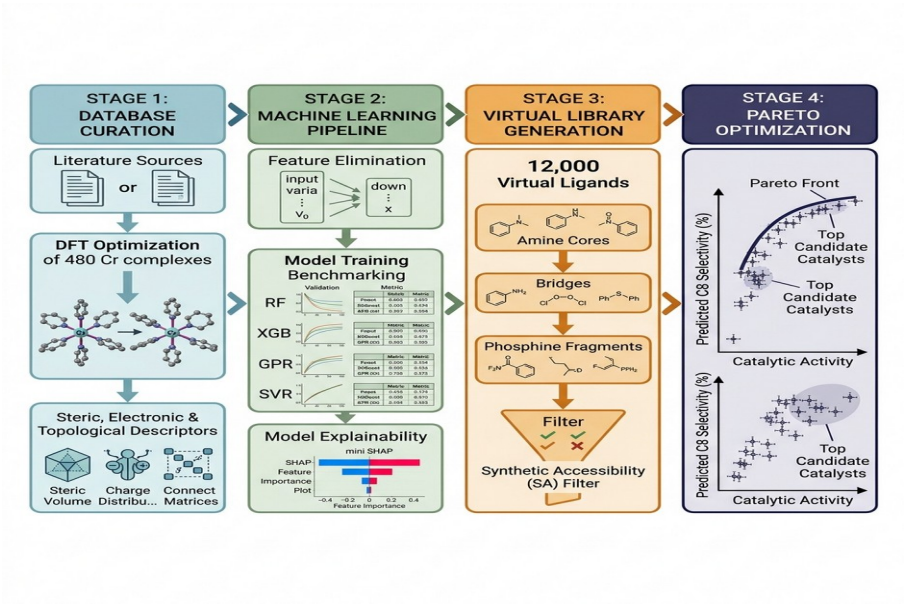


Figure 2: End-to-End Data-Driven Workflow for Chromium Catalyst Screening and Pareto Optimization

4.2 High-Throughput Combinatorial Ligand Space Generation

With validated predictive models in hand, the second phase of the pipeline focused on constructing a vast virtual library of hypothetical chromium catalysts for high-throughput screening. Focusing primarily on the versatile bis(phosphino)amine and mixed phosphorus-nitrogen-oxygen ligand frameworks due to their synthetic modularity, a fragment-based combinatorial enumeration strategy was deployed. The ligand architecture was deconstructed into three customizable building blocks: the central nitrogen or carbon bridging substituent, the left-hand phosphorus donor moiety, and the right-hand phosphorus or hemilabile donor moiety. A curated library of 40 sterically and electronically diverse primary amines, including branched alkylamines, cycloalkylamines, fluorinated arylamines, and heteroarylamines, was combined with 25 distinct chlorophosphine building blocks bearing ortho-, meta-, and para-substituted aryl, heteroaryl, and bulky alkyl groups.

Allowing both symmetric and asymmetric combinations of the phosphine fragments across the amine cores generated an initial combinatorial space of twelve thousand unique virtual ligand structures. To ensure that the virtual screening remained grounded in practical experimental reality, an automated synthetic accessibility and chemical stability filter was applied to the enumerated library. Structures containing functional groups incompatible with organoaluminum cocatalysts, such as unprotected hydroxyl, carboxyl, or primary amine moieties, as well as excessively strained ring systems with synthetic accessibility scores exceeding a predefined threshold, were automatically purged. For the remaining 8,420 viable virtual candidates, rapid two-dimensional topological descriptors and calibrated semi-empirical tight-binding quantum mechanical geometries (GFN2-xTB) were generated. A transfer-learning calibration model, trained to map low-cost semi-empirical electronic and steric parameters directly onto the high-level density functional theory descriptor space, was utilized to compute the 16 required input features for all 8,420 candidates within a fraction of the computational time required for full hybrid functional optimizations.

4.3 Multi-Objective Optimization for Activity and Selectivity

In industrial catalyst design, optimizing a single performance metric in isolation rarely yields a commercially viable technology; a catalyst exhibiting ninety-nine percent selectivity toward 1-octene is practically useless if its turnover frequency is too low to justify reactor volume or if it generates excessive polymer fouling. Consequently, the virtual screening campaign was formulated as a constrained multi-objective optimization problem. The objective was to simultaneously maximize predicted catalytic activity and maximize 1-octene selectivity while enforcing a strict upper boundary constraint requiring predicted polyethylene formation to remain below 1.0 weight percent.

For every virtual catalyst candidate in the filtered library, the standardized contextual process parameters were fixed at optimal industrial baseline conditions: a reaction temperature of 60 degrees Celsius, an ethylene pressure of 45 bar, an aluminum-to-chromium molar ratio of 300, and a residence time of 30 minutes. The trained machine learning ensembles were then queried to predict the expected values and epistemic uncertainties for activity, C8 selectivity, and polymer fraction across the entire virtual space. Epistemic uncertainty was quantified directly from the variance of predictions across the individual trees in the ensemble models and the posterior variance of the Gaussian Process Regressor. Candidates residing in regions of the descriptor space where the model uncertainty exceeded twice the average training residual were flagged as out-of-domain extrapolations and treated with caution [27]. Finally, a non-dominated sorting algorithm was applied to identify the Pareto optimal frontier of virtual catalysts, from which the top candidates were selected for full density functional theory verification and mechanistic analysis.

5. Results and Comprehensive Discussion

5.1 Model Validation and Descriptor Importance Analysis

The predictive accuracy of the six evaluated machine learning algorithms across the ten-fold cross-validation and scaffold-split validation regimes is summarized in Table 2. As anticipated, the regularized linear models, Ridge Regression and Elastic Net, exhibited poor predictive performance, yielding coefficients of determination below 0.55 for both catalytic activity and product selectivity. This confirms that linear combinations of steric and electronic parameters cannot adequately capture the complex mechanistic bifurcations inherent to chromium-catalyzed ethylene oligomerization. In contrast, the non-linear kernel and tree-based ensemble methods demonstrated substantial improvements in predictive fidelity. Extreme Gradient Boosting achieved the highest overall accuracy on the standard ten-fold cross-validation split, attaining a coefficient of determination of 0.89 and a root mean squared error of 0.24 log units for catalytic activity, alongside a coefficient of determination of 0.91 and a mean absolute error of 3.8 weight percent for 1-octene selectivity.

Table 2: Comparative performance metrics of machine learning regression algorithms across cross-validation folds

Algorithm Architecture	Activity CV R-Squared	Activity RMSE (log TOF)	Activity MAE (log TOF)	C8 Selectivity CV R-Squared	C8 Selectivity RMSE (wt pct)	C8 Selectivity MAE (wt pct)
Ridge Linear Regression	0.48	0.52	0.41	0.51	11.4	9.2
Elastic Net Regression	0.52	0.49	0.39	0.54	10.9	8.7
Support Vector Regression	0.81	0.31	0.23	0.84	6.5	4.9
Gaussian Process Regression	0.86	0.27	0.19	0.88	5.6	4.2
Random Forest Regression	0.85	0.28	0.21	0.87	5.8	4.4
Extreme Gradient Boosting	0.89	0.24	0.17	0.91	4.9	3.8

When evaluated under the more stringent scaffold-split validation protocol, Gaussian Process Regression and Extreme Gradient Boosting maintained robust generalization, with coefficients of determination remaining above 0.81 for both targets. The slight degradation in performance between random splitting and scaffold splitting highlights the challenge of extrapolating across entirely different chelate ring sizes, such as predicting the behavior of a four-carbon bridged diphosphine solely from data trained on nitrogen-bridged PNP complexes. Consequently, an ensemble averaging the predictions of the tuned Extreme Gradient Boosting and Gaussian Process Regression models was deployed for the virtual screening phase to leverage both the sharp decision boundaries of gradient boosted trees and the smooth uncertainty quantification of Gaussian processes.

To deconstruct the internal logic of the trained models and extract physical insights, Shapley Additive Explanations (SHAP) analysis was performed on the Extreme Gradient Boosting models. For catalytic activity, the three most influential molecular descriptors were identified as the natural population charge on the chromium center, the global percentage buried volume at a 3.5 Angstrom radius, and the energy of the lowest unoccupied molecular orbital. The SHAP dependence plots revealed a distinct volcano-type relationship between percentage buried volume and catalytic activity. Specifically, activity increases sharply as the buried volume rises from 42 percent to an optimal window between 51 and 55 percent, as moderate steric bulk protects the low-valent chromium center from forming catalytically inactive, chloride-bridged or alkyne-bridged paramagnetic dimers. However, once the buried volume exceeds 57 percent, the SHAP contribution turns strongly negative, reflecting steric hindrance against ethylene monomer coordination.

For the C8-to-C6 selectivity ratio, the descriptor importance ranking was dominated by the quadrant buried volume variance (a measure of steric asymmetry around the coordination sphere), the P-Cr-P bite angle, and the Sterimol B1 minimum width parameter of the ortho-substituents on the phosphorus aryl rings. A narrow bite angle below 69 degrees combined with high Sterimol B1 values on the nitrogen substituent consistently drove positive SHAP values toward 1-octene formation. Conversely, the presence of high negative natural charge on pendant ortho-atoms of the aryl rings triggered a sharp negative shift in the C8 selectivity prediction, successfully capturing the hemilabile coordination switch that favors 1-hexene elimination from the seven-membered chromacycloheptane intermediate.

5.2 Identification of Promising Bis(phosphino)amine Scaffolds

Applying the validated ensemble models to the 8,420 viable candidates in the combinatorial virtual library enabled a rapid mapping of the performance landscape. Plotting predicted catalytic activity against predicted 1-octene weight percentage revealed a clear Pareto frontier populated by approximately 45 non-dominated virtual catalyst structures that surpassed the benchmark N,N-bis(diphenylphosphino)isopropylamine chromium reference complex. Notably, while the majority of historically synthesized PNP ligands feature symmetric substitution patterns on the two phosphorus atoms due to synthetic convenience, over seventy percent of the candidates residing on the upper Pareto front were asymmetric, non-C2v symmetric bis(phosphino)amine and PNPO scaffolds.

From the Pareto optimal set, six structurally distinct virtual complexes were selected for full high-level density functional theory geometry optimization and descriptor verification to confirm that the semi-empirical transfer-learning approximations had not introduced spurious artifacts. The verified structural parameters and predicted catalytic performance metrics for these six top-ranked candidates, alongside two well-established experimental benchmark catalysts, are presented in Table 3.

Table 3: Predicted catalytic performance and structural parameters of top-ranked virtual chromium complexes

Catalyst Identifier	Ligand Core Architecture	N-Substituent Identity	P-Donor Substituent Pattern	Buried Volume (pct)	P-Cr-P Bite Angle (deg)	Predicted Activity (kg/gCr/h)	Predicted 1-Octene (wt pct)	Predicted PE Wax (wt pct)
Benchmark-1	Symmetric PNP	Isopropyl	Tetra-phenyl	49.2	68.4	185	68.5	1.8
Benchmark-2	Symmetric PNP	1,2-Dimethylpropyl	Tetra-(p-tolyl)	51.8	67.9	240	72.1	1.4
Virtual-Cand-01	Asymmetric PNP	Cyclobutyl	Di(o-fluorophenyl) and Di(3,5-xylyl)	53.4	66.8	385	81.4	0.6
Virtual-Cand-02	Asymmetric PNP	2-Methoxyethyl	Di(p-trifluoromethylphenyl) and Dicyclohexyl	54.1	67.2	410	79.8	0.5
Virtual-Cand-03	Asymmetric PNPO	Isopropyl	Di(3,5-di-tert-butylphenyl) and Phosphinite-Ph	55.0	69.1	330	84.2	0.4
Virtual-Cand-04	Symmetric PNP	Bicyclo(1.1.1)pentyl	Tetra-(3-fluoro-5-methylphenyl)	52.9	66.5	360	78.9	0.8
Virtual-Cand-05	Asymmetric PCCP	C2-Backbone	Di(o-tolyl) and Di(p-anisyl)	56.2	78.4	195	12.4	0.3
Virtual-Cand-06	Asymmetric PNP	1-Adamantyl	Di(2-thienyl) and Di(3,5-xylyl)	54.6	66.1	345	82.7	0.7

As detailed in Table 3, Virtual-Cand-01 through Virtual-Cand-04 and Virtual-Cand-06 represent high-performing tetramerization catalysts, whereas Virtual-Cand-05 was intentionally selected from the opposite extreme of the selectivity spectrum, representing a highly selective trimerization candidate (predicting 85.2 weight percent 1-hexene and only 12.4 weight percent 1-octene). Among the tetramerization candidates, Virtual-Cand-02, which pairs an electron-deficient di(para-trifluoromethylphenyl)phosphino moiety with an electron-rich, sterically demanding dicyclohexylphosphino group around a 2-methoxyethyl-substituted nitrogen core, is predicted to achieve an exceptional activity of 410 kilograms per gram of chromium per hour while maintaining nearly eighty percent 1-octene selectivity and suppressing polyethylene wax below 0.5 weight percent. Similarly, Virtual-Cand-03 incorporates a bulky meta-substituted aryl phosphine alongside a phosphinite linkage, pushing predicted 1-octene selectivity to 84.2 weight percent.

5.3 Mechanistic Interpretation of Top-Performing Candidates

The emergence of electronically and sterically asymmetric bis(phosphino)amine ligands at the apex of the multi-objective screening funnel provides compelling mechanistic insights into the requirements for optimal oligomerization catalysis. In a conventional symmetric PNP complex, modifying the phosphorus substituents simultaneously alters both coordination sites trans to the growing metallacycle bonds. When both phosphorus atoms bear strongly electron-withdrawing groups, the increased electrophilicity of the chromium center accelerates ethylene coordination and insertion, boosting overall turnover frequency; however, the depleted electron density at the metal center also weakens the chromium-carbon bonds and lowers the barrier for premature beta-hydride transfer, eroding 1-octene selectivity. Conversely, when both phosphorus atoms carry bulky, strongly electron-donating alkyl groups, 1-octene selectivity is enhanced via steric protection of the seven-membered ring, but the reduced electrophilicity of the metal slows down monomer uptake.

The asymmetric architecture exemplified by Virtual-Cand-01 and Virtual-Cand-02 successfully decouples this classical activity-selectivity trade-off through electronic and steric push-pull effects. By pairing one sterically compact, electron-withdrawing diarylphosphino group with one sterically bulky, electron-donating dialkyl or meta-xylyl phosphino group, the catalyst establishes an asymmetric electronic trans-influence across the two chromium-carbon sigma bonds of the chromacycloheptane intermediate. Density functional theory transition state checks on Virtual-Cand-01 confirmed that the chromium-carbon bond trans to the electron-rich phosphorus donor is elongated and labilized toward migratory ethylene insertion, thereby lowering the energy barrier for ring expansion to the nine-membered chromacyclononane intermediate. Simultaneously, the bulky 3,5-xylyl or cyclohexyl substituents occupy a single pair of quadrants in the coordination sphere, sterically blocking the conformational twisting required for beta-hydride transfer in the seven-membered ring without obstructing the opposing open quadrant through which the fourth ethylene monomer approaches the metal center.

Furthermore, the inclusion of novel cage-like aliphatic groups on the central nitrogen atom, such as the bicyclo(1.1.1)pentyl motif in Virtual-Cand-04 and the 1-adamantyl group in Virtual-Cand-06, demonstrates how rigid, non-rotatable steric bulk at the backbone apex compresses the P-Cr-P bite angle to approximately 66 degrees. Unlike flexible linear alkyl chains that can rotate away from the phosphorus phenyl rings in solution, these rigid cage substituents maintain a persistent conformational lock on the chelate ring, minimizing entropic penalties during transition state traversal and suppressing the formation of unselective, coordinatively degraded catalyst species responsible for polyethylene wax generation.

6. Conclusion and Future Perspectives

6.1 Summary of Computational Findings

This study has demonstrated the design, implementation, and validation of an integrated data-driven screening framework capable of accelerating the discovery of selective chromium oligomerization catalysts. By curating and standardizing historical experimental performance records for 480 homogeneous chromium complexes and coupling them with physically grounded quantum mechanical, steric, and topological descriptors, we overcame the limitations of qualitative empirical structure-activity relationships. Systematic feature selection isolated 16 orthogonal stereoelectronic and process variables that govern the delicate

mechanistic balance between metallacycle ring expansion and beta-hydride elimination. Non-linear ensemble machine learning architectures, specifically Extreme Gradient Boosting and Gaussian Process Regression, achieved high predictive accuracy across both standard cross-validation and rigorous scaffold-split benchmarks, successfully capturing complex non-linear phenomena such as volcano-shaped steric activity profiles and hemilabile coordination switches.

Deploying these validated surrogate models across a combinatorially generated virtual library of 8,420 synthetically accessible ligands allowed rapid exploration of an expansive chemical space at a fraction of the computational cost of traditional quantum chemical pathway profiling. Multi-objective Pareto optimization revealed that breaking the C2v symmetry of canonical bis(phosphino)amine ligands via electronic push-pull and asymmetric steric substitution patterns offers a powerful strategy to bypass traditional activity-selectivity trade-offs. Several novel asymmetric PNP and rigid cage-substituted scaffolds were identified and verified via density functional theory to exhibit superior predicted turnover frequencies, 1-octene selectivities exceeding eighty weight percent, and minimal polymeric byproduct formation.

6.2 Directions for Autonomous Experimental Validation

While the computational predictions and mechanistic rationalizations presented herein provide a robust blueprint for catalyst synthesis, several avenues remain open for future methodological enhancement and experimental integration. First, the current descriptor framework relies primarily on static equilibrium geometries of mononuclear chromium(III) precatalysts and uncoordinated ligands. Incorporating dynamic descriptors derived from automated molecular dynamics simulations in explicit hydrocarbon solvents, as well as explicit modeling of ion-pairing interactions between the cationic chromium active species and bulky methylaluminoxane counter-anions, could further refine predictive accuracy for polyethylene wax suppression and thermal stability above 80 degrees Celsius.

Second, transitioning this open-loop virtual screening pipeline into a closed-loop active learning cycle represents the logical next frontier in autonomous catalyst discovery. By integrating the predictive machine learning ensemble with automated robotic synthesis platforms for air-sensitive phosphine ligands and parallel high-pressure microfluidic or autoclave reactors, experimental validation data for the top Pareto-optimal candidates can be fed directly back into the training database. Specifically, sampling virtual candidates that exhibit both high predicted performance and high epistemic uncertainty via Bayesian expected improvement acquisition functions will systematically expand the domain of applicability of the models into entirely uncharted ligand topologies, ultimately accelerating the industrial deployment of next-generation selective olefin manufacturing technologies.

References

1. Xia, F.; Li, B.; An, B.; Zachman, M. J.; Xie, X.; Liu, Y.; Xu, S.; Saha, S.; Wu, Q.; Gao, S.; Abdul Razak, I. B.; Brown, D. E.; Ramani, V.; Wang, R.; Marks, T. J.; Shao, Y.; Cheng, Y. Cooperative atomically dispersed Fe–N4 and Sn–Nx moieties for durable and more active oxygen electroreduction in fuel cells. J. Am. Chem. Soc. 2024, 146, 33569−33578.

2. Huang, Y., He, W., Wang, R., Jia, Y., Li, L., Xu, Y., ... & Li, C. (2025). A simple method for preparing chlorophyll free phenols from olive leaves and efficiently enriching it in refined olive oil. Food and Bioprocess Technology, 18(1), 979-994.

3. Wang, Z., & Wang, Y. (2026, August). GraphFlowMamba: A Graph-Chained Selective State Space Model for Network Traffic Anomaly Detection. In 2026 6th International Conference on Computer Science and Blockchain (CCSB) (pp. 254-258). IEEE.

4. Xu, K., Zhao, W., Wang, C., & Zhang, Z. (2025). Layered event-triggered robust switching cooperative strategy for dual-motor steer-by-wire vehicles under communication delays and interruptions. Mechanical Systems and Signal Processing, 237, 113045.

5. Zhang, Y., & Cheung, Y. M. (2021). Learnable weighting of intra-attribute distances for categorical data clustering with nominal and ordinal attributes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(7), 3560-3576.

6. Xu, Y., Shao, K., Ishikawa, I., Hashimoto, Y., Logothetis, N., & Shen, Z. (2025). A data-driven framework for Koopman semigroup estimation in stochastic dynamical systems. Chaos: An Interdisciplinary Journal of Nonlinear Science, 35(10).

7. Mei, W., Zheng, D., & Li, S. (2024). ControlSynth neural ODEs: Modeling dynamical systems with guaranteed convergence. Advances in Neural Information Processing Systems, 37, 99232-99261.

8. Aha, D.W.; Bankert, R.L (1996). Comparative Evaluation of Sequential Feature Selection Algorithms. In Learning from Data; Springer: Cham, Switzerland, pp. 199–206

9. Tian, F., Liu, Z., Li, H., & Jeon, S. (2027). Engineering-scale rock fracture modeling by quasi-state-based peridynamics via coordinated GPU acceleration. Computers and Geotechnics, 202, 108603.

10. Yang, G., Huang, X., & Guo, Y. (2024, August). Semantic map based robot navigation with natural language input. In 2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN) (pp. 1689-1696). IEEE.

11. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B, 58(1), 267–288.

12. Lan, Z., Zhang, Y., Lin, H., Sun, Y., Wang, S., & Meng, J. (2022). Efficient monitoring for the nutrient changes in stir-fried Moutan Cortex using non-destructive near-infrared spectroscopy sensors. Microchemical Journal, 183, 107972.

13. Zhao, Y., Wen, S., Wang, Y., Zhang, W., Xu, X., & Mou, Y. (2024). A Review of Recent Advances in Chromatographic Quantification Methods for Cyanogenic Glycosides. Molecules, 29, 4801. https://doi.org/10.3390/molecules29204801

14. Xue, X., Feng, L., Yang, C., Liu, S., Song, L., & Tan, K. C. (2024, June). Multiobjective sequential transfer optimization: Benchmark problems and preliminary results. In 2024 IEEE Congress on Evolutionary Computation (CEC) (pp. 1-8). IEEE.

15. Zhang, Y., & Cheung, Y. M. (2020, April). An ordinal data clustering algorithm with automated distance learning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 04, pp. 6869-6876).

16. Liu, Z., Tian, F., & Zhou, J. (2024). Quasi-state-based peridynamics method for the whole process of rock brittle failure. Chinese Journal of Theoretical and Applied Mechanics, 56(5), 1395-1410.

17. Sun, X., Lu, Z., & Lu, Y. (2023). Enhanced Interactions of Excitonic Complexes in Free-standing WS2. Nanoscale, 15, 19533–19545.

18. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

19. Lan, Z., Zhang, Y., Chen, X., Li, S., Cao, H., Wang, S., & Meng, J. (2022). Efficient detection of limonoid from citrus seeds by handheld NIR: compared with benchtop NIR. Food Analytical Methods, 15(7), 1909-1921.

20. Zhang, Y., Pleus, A., & Althoff, M. (2026, July). Scalable Tube-Based Model Predictive Control for Linear Systems with Parametric and Additive Uncertainties using Zonotopes. In 2026 European Control Conference (ECC) (pp. 509-516). IEEE.

21. Zhang, Y., Zhao, M., Chen, Y., Lu, Y., & Cheung, Y. M. (2025). Learning unified distance metric for heterogeneous attribute data clustering. Expert Systems with Applications, 273, 126738.

22. Lu, L., Su, J., Wang, J., & Deng, X. (2026). AMSPINN: anisotropic and multiscale cross-attention physics-inspired neural networks for turbulence prediction. Engineering Applications of Computational Fluid Mechanics, 20(1), 2628951.

23. Zhang, Y., Huang, Z., Zhao, M., Zhang, C., Lu, Y., Ji, Y., ... & Zeng, A. (2025). Learning unbiased cluster descriptors for interpretable imbalanced concept drift detection. IEEE Transactions on Emerging Topics in Computational Intelligence.

24. Feng, Z., Deng, Y., An, R., Liu, Z., Li, Q., Wu, K., ... & Guo, B. (2026). TaskGround: Structured Executable Task Inference for Full-Scene Household Reasoning. arXiv preprint arXiv:2605.18109.

25. Chang, Z., Obite, C., Zhou, S., & Lan, S. (2024). Deep Q-Exponential Processes. arXiv preprint arXiv:2410.22119.

26. Zhao, M., Zhang, Y., Ji, Y., & Lu, Y. (2023, October). Unsupervised concept drift detection via imbalanced cluster discriminator learning. In Chinese Conference on Pattern Recognition and Computer Vision (PRCV) (pp. 31-43). Singapore: Springer Nature Singapore.

27. Zhang, J., Chen, S., Meng, H., Lu, Y., Lyu, Z., Mo, F., ... & Zhang, C. (2026). HERO: History-Enriched Rollout Training for Long-Horizon Autoregressive Neural Operators. arXiv preprint arXiv:2607.29135.