PANGAEA: AGENTIC SEARCH OVER A UNIFIED CONSTRUCTION SPACE FOR AUTOMATED MOLECULAR THREE-DIMENSIONAL DESCRIPTOR DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Chemically informed molecular descriptors remain valuable for machine-learning based molecular property prediction, especially in data-limited regimes. Among these, three-dimensional (3D) descriptors are particularly appealing due to their ability to encode rich spatial chemical information. However, task-specific 3D descriptor design requires substantial expertise and remains difficult to automate. We abstract 3D descriptor construction as a three-stage process consisting of Formation, Extraction, and Compression, starting from chemically annotated molecular point-cloud sources. This abstraction enables new descriptors to be constructed by combining stage-wise operators and their parameters, thereby reformulating task-specific 3D descriptor design as a search for the combination best suited to each task. Building on this formulation, we propose *Pangaea*, which provides a Unified Construction Space that structures candidate descriptors as executable pipelines, together with an Incumbent Revision Agent that uses a large language model to iteratively refine the best-performing pipeline based on evaluation history, design-space semantics and chemical priors. QM9 experiments under data-limited conditions show *Pangaea* ranks first among descriptor-based methods, achieves a 24% relative reduction in test MAE, and outperforms four deep-learning baselines.

1 INTRODUCTION

Molecular property prediction is a fundamental problem in chemistry, which is important for drug discovery and materials design (Wang et al., 2025; Gao et al., 2025). However, in specialized areas such as formulation design (Samanta et al., 2026) and reaction-condition assessment (Atz et al., 2024), only limited data are typically available for modeling (He et al., 2024). Under these conditions, descriptor-based machine learning (ML) methods remain a competitive choice (Deng et al., 2023). To achieve better predictive performance, descriptor design offers an important avenue beyond model refinement by encoding more effective chemical priors into the features (see Section 3). Molecular descriptors can represent different aspects of molecules, from composition and bond topology to three-dimensional (3D) information (Todeschini & Consonni, 2009). Recent studies suggest that incorporating 3D descriptors can provide complementary spatial information and improve prediction for molecular properties (Campbell & Vogiatzis, 2025; Bahia et al., 2023).

However, designing molecular 3D descriptors for a given task remains challenging, as navigating the vast space of possible spatial representations requires substantial chemical expertise (Pinheiro et al., 2021; Musil et al., 2021). Moreover, descriptor design is often task-specific, requiring repeated expert effort when adapting to new tasks (Faber et al., 2017). Existing automated methods primarily apply symbolic regression to predefined molecular features (Ouyang et al., 2018) or perform parameter tuning on existing 3D descriptors (Barnard et al., 2023), while the systematic design of representations for molecular 3D information remains relatively underexplored. A central question therefore remains: **How can we efficiently design task-specific molecular 3D descriptors?**

To address this question, we examine 20 common classes of 3D descriptors (see Appendix A) and uncover two recurring patterns, illustrated in Figure 1. *First, their inputs can be represented as chemically annotated point clouds in three-dimensional space. Second, their construction follows a three-stage pipeline: selecting a point cloud to form a 3D carrier, extracting an intermediate*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

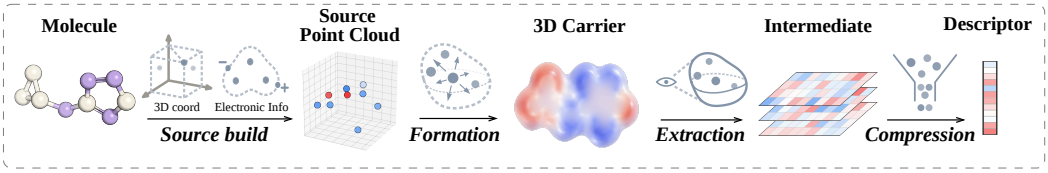


Figure 1: **Shared construction of 3D descriptors.** For each molecule, source point clouds are generated from molecular geometry and electronic structure. Each Formation operator selects its associated source and forms a carrier, which Extraction and Compression transform into a descriptor.

representation from the carrier, and compressing this representation into a descriptor. This decomposition enables new 3D descriptors to be constructed by recombining parameterized Formation, Extraction, and Compression operators. We therefore formulate task-specific 3D descriptor design as a search for the combination of stage-wise operators and parameters best suited to each task.

In this paper, we propose *Pangaea*, a framework for task-specific 3D descriptor design. It provides a Unified Construction Space (UCS), a design space derived from existing 3D descriptors and organized into four branches, with dimensions specifying operator choices and parameters, spanning the formation of 3D carrier from selected source and its encoding. To better explore the UCS, *Pangaea* employs an Incumbent Revision Agent (IRA), a large language model (LLM)-based agent that iteratively revises the incumbent, defined as the best pipeline found so far, based on evaluation history, design-space semantics describing what each UCS dimension controls, and chemical priors. Experiments on a QM9 subset emulating a data-limited regime show that *Pangaea* ranks first among descriptor-based methods under each of three ML learners, and its best result across the learners outperforms the best DL baseline on 11 of 12 tasks. Meanwhile, on the Fe Mössbauer benchmark, *Pangaea* further reduces prediction error relative to expert-crafted descriptors that encode spatial information about the coordination environment. In summary, our major contributions are as follows:

(1) We provide a general formulation for task-specific molecular 3D descriptor design to the cheminformatics community by abstracting descriptor construction as a three-stage process. This formulation enables descriptors to be designed by combining stage-wise operators and parameters, reframing molecular 3D descriptor design as a search for the combination best suited to each task.

(2) We propose *Pangaea*, a framework for task-specific 3D descriptor design, with a Unified Construction Space that represents candidate pipelines in terms of operator choices and associated parameters, alongside an Incumbent Revision Agent that iteratively searches the space for descriptors suited to each task using evaluation history, design-space semantics, and chemical priors.

(3) We evaluate *Pangaea* on QM9 in a data-limited setting. It ranks first among descriptor-based methods for each of three ML learners, with a 24% average relative reduction in test MAE compared with using graph and global descriptors alone, while surpassing DL baselines on 11 of 12 tasks.

2 Related Work

Automated descriptor design. Automated descriptor design has been studied in molecular modeling. Existing approaches focus on applying symbolic regression to molecular features (Ouyang et al., 2018; Hernandez et al., 2019) or tuning parameters of predefined 3D descriptors (Barnard et al., 2023). However, automating the process of transforming 3D molecular information into task-specific descriptors remains underexplored. *Pangaea* addresses this gap by decomposing 3D descriptor construction into composable stages, thereby recasting automated 3D descriptor design as identifying the combination of stage-wise methods and parameters best suited to each task.

LLM-guided descriptor design. In recent years, LLMs have increasingly been explored for descriptor design. However, prior work has primarily used LLMs to select established descriptors (Gómez-Tamayo et al., 2025) or to propose and refine features from molecular topology and properties using chemical knowledge (Zheng et al., 2025; Ning et al., 2026; Cobelli & Sanvito, 2026), with little attention to how LLMs can guide the encoding of 3D spatial information. In contrast, *Pangaea* couples an LLM agent with a Unified Construction Space, enabling structured reasoning over alternative ways of encoding 3D spatial information.

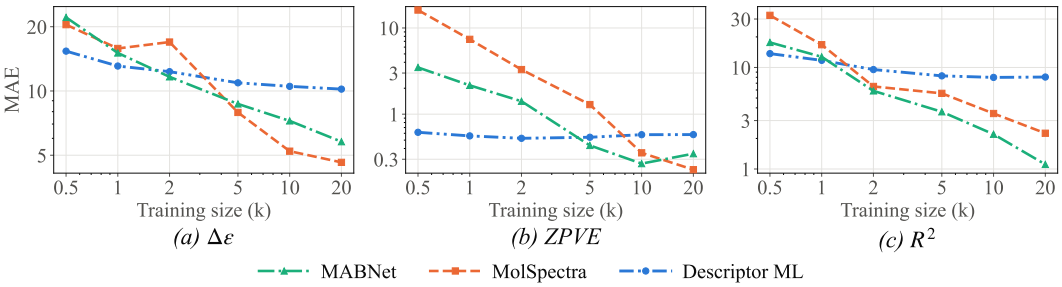


Figure 2: **Data-scaling behavior for molecular property prediction.** Test mean absolute error (MAE) on the same test set across three QM9 tasks as the training-set size increases, for MABNet (supervised), MolSpectra (pretrained and fine-tuned) and Descriptor ML.

3 Motivation

Molecular property prediction with limited data. Recent end-to-end deep learning (DL) methods learn molecular representations directly, avoiding hand-crafted descriptors (Schütt et al., 2017; Wang et al., 2024), but often fail to outperform ML methods when data are limited (Jiang et al., 2021; Sultan et al., 2026). To examine this, we compare two state-of-the-art DL methods, MABNet (Rao et al., 2025) and MolSpectra (Wang et al., 2025), with an ML method using both topological and 3D descriptors following common practice (Campbell & Vogiatzis, 2025; Bahia et al., 2023) across training sizes of 0.5k–20k molecules, with all methods evaluated on the same test set. As shown in Figure 2, with 500 training molecules, the DL methods have 1.3–26 times the prediction error of the ML model and require several thousand training molecules to close the gap. Pretraining, as used in MolSpectra, can improve accuracy but does not eliminate dependence on downstream data (Tripuraneni et al., 2020; Du et al., 2021). Therefore, ML methods remain effective in the low-data regime, but depend on appropriate molecular descriptors (Deng et al., 2023). In practice, designing such descriptors is often expertise-intensive, especially for 3D representations (Pinheiro et al., 2021; Musil et al., 2021), motivating automated methods for efficient 3D descriptor design.

Organizing candidate descriptor constructions. The shared Formation, Extraction, and Compression pipeline provides a common structure for composing 3D descriptors. To organize the possible combinations, we further examine existing 3D descriptors. As shown in Table S1, their construction approaches exhibit four broad patterns based on how they represent chemical signals and process geometry. Specifically, descriptors such as the Coulomb matrix (Rupp et al., 2012) encode geometric relations between points, while CoMFA (Cramer et al., 1988) constructs and samples chemical fields. Methods such as WHIM (Todeschini et al., 1994) and SOAP (Bartók et al., 2013) represent geometry through basis expansions. HDAD (Faber et al., 2017) aggregates geometric quantities into statistical profiles. Guided by these patterns, we organize candidate constructions in the UCS into four branches: *point*, *field*, *expansion*, and *statistics*, each following the shared three-stage pipeline.

Effective search over descriptor pipelines. Black-box optimization is widely used to identify high-performing configurations through iterative evaluations (Larson et al., 2019). However, its search efficiency can deteriorate as the dimensionality and complexity of the search space increase (Rios & Sahinidis, 2013). To examine search behaviors within the UCS, we visualize the incumbent trajectories of different strategies, tracing how their best-found pipelines evolve. As shown in Figure 3 (a), two representative black-box optimizers, TPE (Bergstra et al., 2011) and PSO (Kennedy & Eberhart, 1995), improve upon their initial

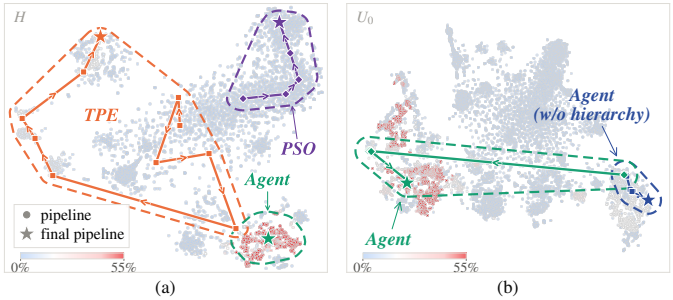


Figure 3: **Incumbent trajectories of different search strategies.** Each point represents a pipeline in a t-SNE embedding, with redder colors indicating better performance.

configurations but remain in suboptimal regions within the evaluation budget. This behavior may partly stem from how these optimizers infer search directions from past evaluations. Specifically, they use evaluation feedback to propose new configurations through predefined statistical or heuristic rules, without explicitly reasoning over the semantics of UCS dimensions. In *Pangaea*, we therefore use an LLM agent to revise the best-found pipeline based on evaluation history, guided by these design-space semantics and task-relevant chemical priors. However, as illustrated in Figure 3 (b), an LLM agent under unconstrained exploration (*Agent w/o hierarchy*) may derive most of its improvements from local refinements around an initially promising pipeline, leaving broader regions of the search space underexplored. We therefore structure the agent's exploration hierarchically, ensuring that both local refinements and broader structural changes are considered throughout the search.

4 Problem Formulation

We study the problem of designing 3D molecular descriptors for molecular property prediction tasks. Let $\mathcal{D} = \{(\mathcal{M}_i, y_i)\}_{i=1}^{N}$ denote a molecular dataset, where $\mathcal{M}_i$ denotes a molecular record containing atomic identities, 3D coordinates, and electronic-structure data, while $y_i \in \mathcal{Y}$ is the property label. Our goal is to find an optimal construction pipeline $\varphi^\star$ in a construction space Φ such that $\varphi^\star = \arg\max_{\varphi\in\Phi} \mathrm{Score}(\varphi; \mathcal{D})$, where $\mathrm{Score}(\varphi; \mathcal{D})$ measures the predictive performance of an ML learner using the baseline descriptors augmented with $\varphi(\mathcal{M})$ (higher is better).

Each candidate pipeline $\varphi \in \Phi$ maps a molecular record $\mathcal{M}$ to a descriptor vector $D = \varphi(\mathcal{M}) \in \mathbb{R}^{K_\varphi}$ through three compatible operators:

$$\varphi(\mathcal{M}): \quad \mathcal{S}(\mathcal{M}) \xrightarrow{\kappa} C \xrightarrow{\tau} I \xrightarrow{\psi} D, \tag{1}$$

where $\mathcal{S}(\mathcal{M})$ is a library of attributed point clouds constructed from the molecule's 3D coordinates and electronic-structure information. The Formation operator κ selects its associated source and forms a carrier C that holds its chemical signal on a spatial support. The Extraction operator τ derives an indexed intermediate representation I from C. The Compression operator ψ encodes I into the descriptor D, whose dimension K_φ is fixed across molecules.

5 Method

5.1 Overview

Figure 4 shows the architecture of *Pangaea*, which comprises two core components: a Unified Construction Space (UCS) and an Incumbent Revision Agent (IRA). Together with a descriptor-evaluation procedure, these components form a closed loop for task-specific 3D descriptor design. In each round, IRA generates a batch of candidate pipeline configurations through multi-level revisions of the incumbent, i.e., the best-scoring candidate found so far, guided by evaluation history queried through analysis tools, design-space semantics, and chemical priors. Each candidate is then instantiated by the UCS as an executable pipeline, whose resulting descriptor is concatenated with the fixed, task-specific baseline descriptors and fed into the ML learner. The predictive performance yields the search score, which is recorded in the evaluation history together with the corresponding pipeline and the reasoning behind its proposal, to inform subsequent iterations. Once the evaluation budget is exhausted, the learner is refitted using the baseline descriptors augmented with descriptors from the best-scoring pipeline, and evaluated on a held-out test split never used during search.

5.2 Unified Construction Space for 3D Descriptors

The Unified Construction Space (UCS) provides a structured organization of candidate 3D descriptor constructions. Each configuration in the UCS corresponds to a descriptor pipeline specified by a combination of Formation, Extraction, and Compression operators, together with their parameters.

Chemical sources and construction operators. UCS assembles a shared source library $\mathcal{S}(\mathcal{M})$ of n_{s} chemical point clouds encoding diverse chemical signals, including atomic identities, electronegativities, charges, and information derived from electronic structure. All source clouds are expressed in a shared canonical frame intrinsic to each molecule, removing dependence of the resulting representation on a molecule's global orientation. In addition, UCS provides stage-wise operator pools

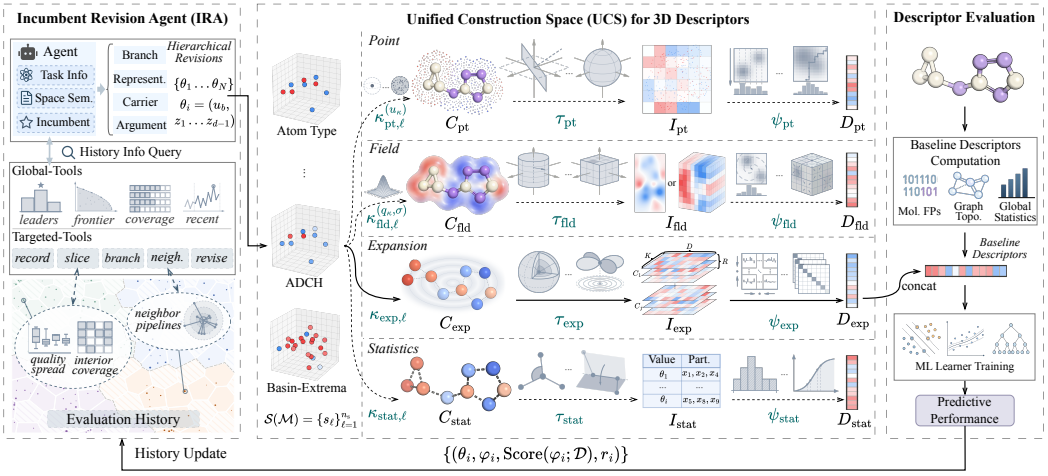


Figure 4: **Overall architecture of *Pangaea*.** The framework comprises two components: (1) a Unified Construction Space (UCS) with four branches organizing candidate pipelines that construct descriptors from a shared chemical source library via the Formation, Extraction, and Compression stages; and (2) an Incumbent Revision Agent (IRA) that uses evaluation history, design-space semantics, and task-relevant chemical priors to guide the search over UCS. A descriptor-evaluation procedure assesses the generated descriptors, linking the two components in a closed loop.

for Formation κ, Extraction τ, and Compression ψ. The three pools draw on operations abstracted from the surveyed descriptors and stage-specific extensions, with each Formation operator selecting its associated source cloud from $\mathcal{S}(\mathcal{M})$. Details are provided in Appendix H. All operators are equipped with explicit input and output contracts to support compatible composition across stages.

Construction space. Following the construction patterns discussed in Section 3, UCS organizes candidate pipelines into four branches: *point*, *field*, *expansion*, and *statistics*, indexed by $\mathcal{B} = \{\mathrm{pt}, \mathrm{fld}, \mathrm{exp}, \mathrm{stat}\}$. For each branch $b \in \mathcal{B}$, UCS filters the shared operator pools to retain Formation, Extraction, and Compression operators applicable to b, yielding three operator sets denoted by $\mathcal{K}_b$, $\mathcal{T}_b$, and $\mathcal{P}_b$, respectively. Candidate pipelines in branch b are constructed by combining operators from each of these sets, together with their corresponding parameters. However, not every such combination yields an executable pipeline. For example, a voxel array obtained by sampling a chemical field can be compressed by voxel moments or voxel spectrum, but not by row pooling, which requires a two-dimensional matrix input. Thus, a pipeline is valid precisely when consecutive operators have compatible output and input contracts.

Accordingly, with $\varphi_{\kappa,\tau,\psi}$ denoting a pipeline in Eq. (1), UCS can be represented as the set of all compatible pipelines across the four branches:

$$\Phi := \bigcup_{b\in\mathcal{B}} \left\{ \varphi_{\kappa,\tau,\psi} \mid (\kappa,\tau,\psi) \in \mathcal{K}_b \times \mathcal{T}_b \times \mathcal{P}_b,\ \mathcal{V}(\kappa,\tau,\psi) \right\}, \qquad \mathcal{V}(\kappa,\tau,\psi) \iff (\mathcal{O}_\kappa \preceq \mathcal{I}_\tau) \wedge (\mathcal{O}_\tau \preceq \mathcal{I}_\psi). \tag{2}$$

Here $\mathcal{I}_\omega$ and $\mathcal{O}_\omega$ denote the input and output contracts of an operator ω, respectively. The relation $\preceq$ indicates that the output contract satisfies the next operator's input requirements, including shape, axes, ordering, channels, and auxiliary information. Appendix H provides the full specification of UCS, including its chemical sources, operator pools, and parameter settings.

A common search interface. UCS is a heterogeneous and conditional space in which different branches involve different operators and parameters, while valid choices depend on preceding selections. Direct search over this space requires explicit handling of branch-specific variables and compatibility constraints. To provide a common interface for different search strategies, *Pangaea* equips UCS with a normalized address space

$$\Theta = [0,1]^d, \qquad \theta = (u_b, z_1, \ldots, z_{d-1}) \in \Theta, \tag{3}$$

where each address is mapped to a candidate pipeline φ_θ by a conditional decoder $\mathrm{Dec} : \Theta \to \Phi$. Here u_b selects one of the four branches by partitioning $[0,1]$ into four equal intervals, while

$z_1, \ldots, z_{d-1}$ specify stage-wise operator choices and their parameters. The mapping from address coordinates to descriptor construction pipelines is detailed in Appendix H.

5.3 Incumbent Revision Agent

To identify a pipeline suited to a given task within the UCS, *Pangaea* employs an Incumbent Revision Agent (IRA) that iteratively refines the incumbent pipeline. At each round, IRA proposes a hierarchically organized batch of revisions to the best-scoring pipeline found so far, guided by evaluation history, design-space semantics and chemical priors.

Proposing revisions. IRA uses an LLM to analyze three sources of information. Specifically, the evaluation history records each evaluated pipeline φ_i, its address θ_i, search score $\mathrm{Score}(\varphi_i; \mathcal{D})$, and associated rationale r_i. IRA uses predefined retrieval tools to obtain relevant information because directly including the full history in the LLM context becomes increasingly inefficient as the history grows. IRA first uses four global tools to summarize pipeline performance, search coverage, recent progress, and opportunities for further exploration. It may then invoke up to two targeted tools to retrieve detailed records and statistics relevant to its current analysis. Details of the retrieval tools are provided in Appendix B.1. Beyond the retrieved history, IRA is also provided with a description of the design space, specifying what each dimension controls, together with a task description that provides chemical prior knowledge. Based on these inputs, IRA formulates revision rationales and translates them into proposed revisions to the incumbent.

Hierarchical revisions. To balance local refinement with broader exploration, IRA allocates each round's proposal budget across four revision levels with progressively narrower scopes. All revisions start from the same incumbent. At the *branch level*, IRA switches to another branch and reconfigures the pipeline. At the *representation level*, it retains the incumbent's branch and revises how the 3D information is represented, primarily through the Extraction and Compression operators, while permitting coordinated changes to other choices and parameters. At the *carrier level*, it retains the incumbent's branch and representation-level choices while focusing on carrier-related changes, mainly through Formation choices and parameters. At the *argument level*, it retunes only continuous parameters to make local adjustments. Details of each level are provided in Appendix B.2.

6 Results

6.1 Evaluation setting

Datasets. To evaluate *Pangaea* under data-limited conditions, we use two molecular property prediction benchmarks. Specifically, we construct *QM9-sample*, a 1498-molecule subset of QM9 (Ramakrishnan et al., 2014), using property-stratified sampling to emulate a small-data setting while aiming to preserve the distribution of the full dataset. We further consider *MB*, a small Fe Mössbauer spectroscopy dataset, which is relevant to understanding iron complexes (Kuzmann et al., 2021), to represent a naturally data-limited task. Appendix C provides details on dataset construction.

Baselines. On *QM9-sample*, we compare 3D descriptors designed by *Pangaea* with three representative 3D descriptors (SOAP (Bartók et al., 2013), LMBTR (Huo & Rupp, 2022; Langer et al., 2022) and GRIND (Pastor et al., 2000)) as augmentations to the fixed *Topology & Global* descriptors, which combine molecular graph descriptors and global physicochemical features, using SVR (Drucker et al., 1996), Ridge (Hoerl & Kennard, 1970) and XGBoost (Chen & Guestrin, 2016) as ML learners. We further compare with four end-to-end DL methods (SchNet (Schütt et al., 2017), ViSNet (Wang et al., 2024), MolSpectra (Wang et al., 2025) and MABNet (Rao et al., 2025)), each retrained under its own training paradigm and evaluated on the same held-out test set. On *MB*, we use *Coordination & Global* as baseline, comprising expert-crafted descriptors that encode spatial information about the local coordination environment. Further details are provided in Appendix D.

Implementation details. In *Pangaea*, we construct a set of point-cloud sources $\mathcal{S}(\mathcal{M})$ from 13 representative source types for each molecule, spanning molecular structure, atomic properties, and electronic information derived from DFT calculations. The address dimension d of UCS is set to 15, the minimum needed to assign a coordinate to every operator choice and parameter across all branches. IRA is powered by `gpt-5.6-terra`. Each search begins with a random warm-up and uses an evaluation budget of 2000 per *QM9-sample* task or 500 per *MB* task. We report mean abso-

Table 1: Performance comparison across the twelve *QM9-sample* tasks. For each task, the **best overall** value is in bold, the best DL value is shaded blue, and the best value within each ML block is shaded red. Red and blue arrows indicate relative improvements and regressions against the *Topology & Global* baseline within each ML block. *Rank* columns (lower is better) report mean ranks over twelve tasks, within each block (*block*) and across all settings (*all*).

Setting	μ	α	$\varepsilon_{\text{HOMO}}$	$\varepsilon_{\text{LUMO}}$	$\Delta\varepsilon$	R^2	ZPVE	U_0	U	H	G	c_v	Rank block	Rank all
SchNet [NIPS'17]	0.515	0.491	22.51	20.13	26.21	10.09	15.71	1050	1049	1050	1050	0.228	3.3	12.9
ViSNet [Nat. Comm.'24]	0.397	**0.298**	10.56	9.977	13.91	99.31	1.375	891	888	887	882	0.137	1.8	10.4
MolSpectra [ICLR'25]	0.222	0.608	17.76	18.51	15.21	13.89	11.49	9101	5379	4098	4103	0.192	3.2	14.2
MABNet [ICML'25]	0.316	0.370	9.008	11.08	16.80	12.28	2.152	683	648	692	688	0.158	**1.7**	10.2
XGBoost														
Topo. & Global	0.664	1.046	8.98	10.32	12.60	20.97	1.059	1753	1723	1723	1746	0.426	3.6	13.7
+SOAP	0.662 $_{\uparrow 0.3\%}$	1.076 $_{\downarrow 2.9\%}$	8.92 $_{\uparrow 0.6\%}$	10.46 $_{\downarrow 1.3\%}$	12.68 $_{\downarrow 0.6\%}$	21.22 $_{\downarrow 1.2\%}$	1.035 $_{\uparrow 2.2\%}$	1749 $_{\uparrow 0.2\%}$	1753 $_{\downarrow 1.7\%}$	1753 $_{\downarrow 1.7\%}$	1760 $_{\downarrow 0.8\%}$	0.422 $_{\uparrow 0.9\%}$	3.8	13.9
+LMBTR	0.662 $_{\uparrow 0.3\%}$	1.019 $_{\uparrow 2.6\%}$	9.11 $_{\downarrow 1.4\%}$	10.25 $_{\uparrow 0.7\%}$	12.59 $_{\uparrow 0.1\%}$	20.76 $_{\uparrow 1.0\%}$	1.016 $_{\uparrow 4.0\%}$	1741 $_{\uparrow 0.7\%}$	1733 $_{\downarrow 0.5\%}$	1733 $_{\downarrow 0.5\%}$	1728 $_{\uparrow 1.0\%}$	0.424 $_{\uparrow 0.6\%}$	2.8	13.0
+GRIND	0.649 $_{\uparrow 2.3\%}$	1.066 $_{\downarrow 2.0\%}$	8.78 $_{\uparrow 2.3\%}$	10.22 $_{\uparrow 1.0\%}$	12.35 $_{\uparrow 2.0\%}$	21.26 $_{\downarrow 1.4\%}$	1.051 $_{\uparrow 0.7\%}$	1813 $_{\downarrow 3.4\%}$	1777 $_{\downarrow 3.1\%}$	1777 $_{\downarrow 3.1\%}$	1778 $_{\downarrow 1.8\%}$	0.420 $_{\uparrow 1.5\%}$	3.6	13.6
+Pangaea	**0.124** $_{\uparrow 81\%}$	0.952 $_{\uparrow 9.0\%}$	8.57 $_{\uparrow 4.6\%}$	9.83 $_{\uparrow 4.7\%}$	**12.21** $_{\uparrow 3.1\%}$	16.23 $_{\uparrow 23\%}$	0.979 $_{\uparrow 7.5\%}$	1738 $_{\uparrow 0.9\%}$	1735 $_{\downarrow 0.7\%}$	1715 $_{\uparrow 0.5\%}$	1671 $_{\uparrow 4.3\%}$	0.419 $_{\uparrow 1.7\%}$	**1.2**	9.9
Ridge														
Topo. & Global	0.769	0.813	9.50	13.24	16.56	17.93	0.610	391	391	391	391	0.194	4.0	10.7
+SOAP	0.744 $_{\uparrow 3.3\%}$	0.760 $_{\uparrow 6.5\%}$	9.34 $_{\uparrow 1.7\%}$	13.25 $_{\downarrow 0.1\%}$	16.66 $_{\downarrow 0.6\%}$	17.59 $_{\uparrow 1.9\%}$	0.547 $_{\uparrow 10\%}$	389 $_{\uparrow 0.7\%}$	389 $_{\uparrow 0.7\%}$	389 $_{\uparrow 0.7\%}$	389 $_{\uparrow 0.7\%}$	0.191 $_{\uparrow 1.1\%}$	2.8	9.5
+LMBTR	0.757 $_{\uparrow 1.6\%}$	0.819 $_{\downarrow 0.7\%}$	9.47 $_{\uparrow 0.3\%}$	13.02 $_{\uparrow 1.7\%}$	16.14 $_{\uparrow 2.5\%}$	17.73 $_{\uparrow 1.1\%}$	0.605 $_{\uparrow 0.8\%}$	387 $_{\uparrow 1.2\%}$	387 $_{\uparrow 1.2\%}$	387 $_{\uparrow 1.2\%}$	387 $_{\uparrow 1.2\%}$	0.194 $_{\downarrow 0.2\%}$	2.8	9.5
+GRIND	0.762 $_{\uparrow 0.9\%}$	0.827 $_{\downarrow 1.7\%}$	9.66 $_{\downarrow 1.6\%}$	13.27 $_{\downarrow 0.2\%}$	16.55 $_{\uparrow 0.1\%}$	17.93 $_{\uparrow 0.1\%}$	0.610 $_{\uparrow 0.1\%}$	396 $_{\downarrow 1.2\%}$	396 $_{\downarrow 1.2\%}$	396 $_{\downarrow 1.2\%}$	396 $_{\downarrow 1.2\%}$	0.189 $_{\uparrow 2.3\%}$	4.4	11.0
+Pangaea	0.144 $_{\uparrow 81\%}$	0.667 $_{\uparrow 18\%}$	9.29 $_{\uparrow 2.3\%}$	12.40 $_{\uparrow 6.4\%}$	15.86 $_{\uparrow 4.3\%}$	8.43 $_{\uparrow 53\%}$	**0.432** $_{\uparrow 29\%}$	265 $_{\uparrow 32\%}$	305 $_{\uparrow 22\%}$	265 $_{\uparrow 32\%}$	**218** $_{\uparrow 44\%}$	0.175 $_{\uparrow 9.6\%}$	**1.0**	5.5
SVR														
Topo. & Global	0.732	0.701	8.16	10.10	13.00	11.18	0.639	460	460	460	460	0.150	3.8	7.9
+SOAP	0.696 $_{\uparrow 5.0\%}$	0.640 $_{\uparrow 8.7\%}$	8.15 $_{\uparrow 0.1\%}$	10.10 $_{\uparrow 0.1\%}$	13.02 $_{\downarrow 0.1\%}$	11.14 $_{\uparrow 0.3\%}$	0.610 $_{\uparrow 4.6\%}$	463 $_{\downarrow 0.6\%}$	463 $_{\downarrow 0.7\%}$	463 $_{\downarrow 0.6\%}$	462 $_{\downarrow 0.4\%}$	0.143 $_{\uparrow 4.3\%}$	3.1	7.0
+LMBTR	0.730 $_{\uparrow 0.3\%}$	0.716 $_{\downarrow 2.2\%}$	**8.10** $_{\uparrow 0.8\%}$	9.94 $_{\uparrow 1.6\%}$	12.92 $_{\uparrow 0.6\%}$	11.16 $_{\uparrow 0.2\%}$	0.653 $_{\downarrow 2.2\%}$	453 $_{\uparrow 1.4\%}$	454 $_{\uparrow 1.4\%}$	453 $_{\uparrow 1.5\%}$	454 $_{\uparrow 1.4\%}$	0.148 $_{\uparrow 1.1\%}$	2.9	7.0
+GRIND	0.724 $_{\uparrow 1.1\%}$	0.709 $_{\downarrow 1.2\%}$	8.17 $_{\downarrow 0.1\%}$	9.93 $_{\uparrow 1.8\%}$	12.93 $_{\uparrow 0.5\%}$	11.43 $_{\downarrow 2.3\%}$	0.629 $_{\uparrow 1.5\%}$	465 $_{\downarrow 1.0\%}$	465 $_{\downarrow 1.1\%}$	465 $_{\downarrow 0.9\%}$	466 $_{\downarrow 1.1\%}$	0.147 $_{\uparrow 1.5\%}$	4.0	8.0
+Pangaea	0.143 $_{\uparrow 80\%}$	0.495 $_{\uparrow 29\%}$	8.23 $_{\downarrow 0.9\%}$	**9.18** $_{\uparrow 9.1\%}$	12.42 $_{\uparrow 4.5\%}$	**6.21** $_{\uparrow 44\%}$	0.460 $_{\uparrow 28\%}$	**247** $_{\uparrow 46\%}$	**248** $_{\uparrow 46\%}$	**252** $_{\uparrow 45\%}$	248 $_{\uparrow 46\%}$	**0.135** $_{\uparrow 9.8\%}$	**1.3**	**2.0**

lute error (MAE) for regression targets and accuracy for classification targets, with results averaged over three random seeds (41–43). Pipeline search and the ML learners run on Intel Xeon Gold CPUs, and the DL baselines on NVIDIA RTX 3090 GPUs. Details can be found in Appendix E.

6.2 Performance in data-limited molecular property prediction

To evaluate *Pangaea* in data-limited molecular property prediction, we compare its designed descriptors with existing 3D descriptors and end-to-end DL methods on *QM9-sample*. As shown in Table 1, augmentation with *Pangaea* yields the best mean rank for each of three ML learners, achieves an overall average relative reduction of 24% in test MAE over the *Topology & Global* baseline, and outperforms its variants with existing 3D descriptors. With SVR, *Pangaea* ranks first among evaluated ML and DL baselines. Across the 12 tasks, the best *Pangaea*-augmented learner surpasses the best DL method on 11, whereas the *Topology & Global* baseline alone does so on 7 tasks and on 8 when combined with existing 3D descriptors. The designed descriptors are detailed in Appendix G.1.

We next evaluate *Pangaea* on the *MB* task, where the *Coordination & Global* baseline descriptors already include spatial information about the coordination environment. As shown in Table 2, augmenting with *Pangaea*-designed descriptors further improves predictive performance across all four tasks and achieves the best mean rank compared with augmentations using existing 3D descriptors. Additional analyses are provided in Appendix G.2.

Table 2: Performance comparison across the four *MB* tasks with XGBoost. Arrows are relative to the *Coordination & Global* baseline. Other notation as in Table 1.

Setting	δ (mm/s ↓)	ΔE_Q (mm/s ↓)	Iron ox. state (% ↑)	Iron spin state (% ↑)	Rank
Coord. & Global	0.1182	0.6417	71.36	57.77	3.5
+SOAP	0.1157 $_{\uparrow 2.0\%}$	0.6563 $_{\downarrow 2.3\%}$	71.84 $_{\uparrow 0.7\%}$	60.19 $_{\uparrow 4.2\%}$	2.8
+LMBTR	0.1138 $_{\uparrow 3.7\%}$	0.6644 $_{\downarrow 3.5\%}$	70.87 $_{\downarrow 0.7\%}$	57.28 $_{\downarrow 0.8\%}$	3.8
+GRIND	0.1209 $_{\downarrow 2.3\%}$	0.6408 $_{\uparrow 0.1\%}$	67.96 $_{\downarrow 4.8\%}$	58.74 $_{\uparrow 1.7\%}$	3.8
+Pangaea	0.1154 $_{\uparrow 2.3\%}$	0.6165 $_{\uparrow 3.9\%}$	73.46 $_{\uparrow 2.9\%}$	61.00 $_{\uparrow 5.6\%}$	**1.3**

6.3 Analysis of the Incumbent Revision Agent

To assess the effectiveness of IRA, we compare it against three black-box optimizers: TPE (Bergstra et al., 2011), CMA-ES (Hansen & Ostermeier, 2001), and PSO (Kennedy & Eberhart, 1995). TPE is run both on normalized pipeline addresses (TPE) and with explicit access to the branch-specific conditional structure (TPE-struct). See Appendix E.4 for more details. As shown in Figure 5, IRA outperforms all three black-box optimizers overall, achieving the best mean rank and the best final

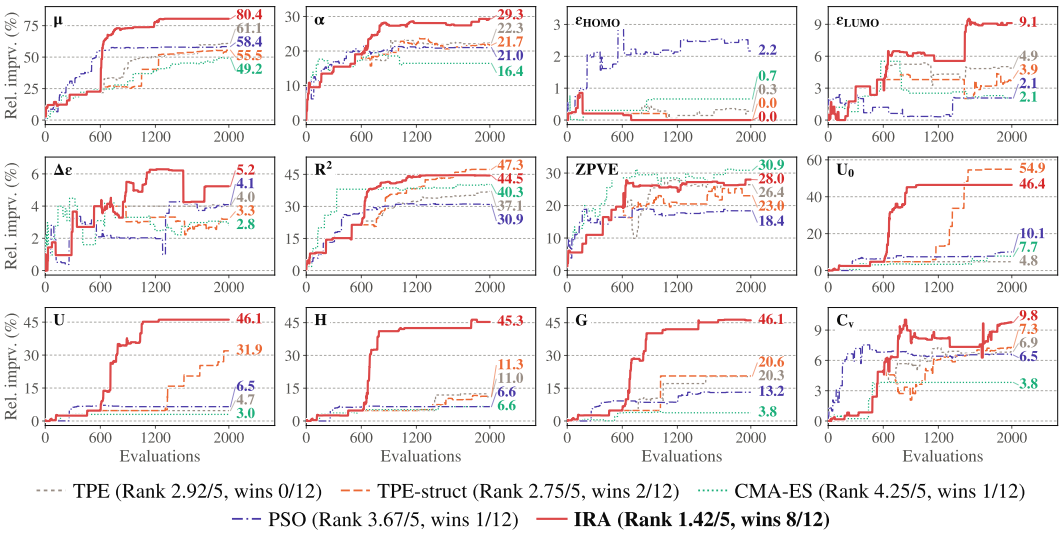


Figure 5: **Comparison of search strategies on *QM9-sample*.** Each curve shows the evolution of the incumbent's held-out test improvement over the *Topology & Global* baseline, assessed after search.

performance on eight of the twelve *QM9-sample* tasks. Across many tasks, IRA exhibits rapid performance improvements and continues to benefit from additional evaluations, while the black-box optimizers show more limited gains. Additional analyses are provided in Appendix F.1.

To quantify the benefit of using design-space semantics and chemical priors to guide search, we construct IRA (w/o guidance) by anonymizing the corresponding information. Figure 6 shows the evolution of the attainment ratio, defined as the incumbent's search-score improvement normalized by the improvement of the best observed pipeline on each task. Starting from the same random warm-up, the full IRA achieves better search performance than the anonymized variant. This performance gap is evident from the early post warm-up evaluations and persists as the evaluation budget increases, indicating that design-space semantics and chemical priors improve both search efficiency and the quality of the discovered descriptors. Additionally, IRA (w/o guidance) still outperforms the black-box optimizers, indicating that IRA retains a search advantage even without semantic or chemical guidance. Details and per-task results are reported in Appendix F.2.

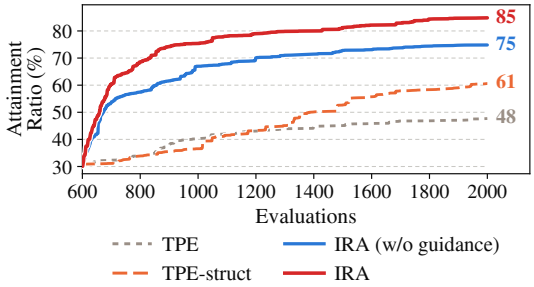


Figure 6: **Analysis of guidance from design-space semantics and chemical priors in IRA.** Curves track the incumbent's mean attainment ratio across the twelve *QM9-sample* tasks.

To assess the impact of hierarchical revisions in IRA, we construct five variants, one removing the hierarchy entirely, allowing the LLM to freely choose what to modify at each search step, and four each omitting one revision level. As shown in Figure 7, removing the hierarchical revision strategy substantially degrades search performance, and ablating any individual revision level also leads to a consistent drop. Among the four levels, removing argument-level revisions has the smallest overall impact, indicating the importance of enabling larger-scale adjustments during the search process. Appendix F.3 presents the detailed results.

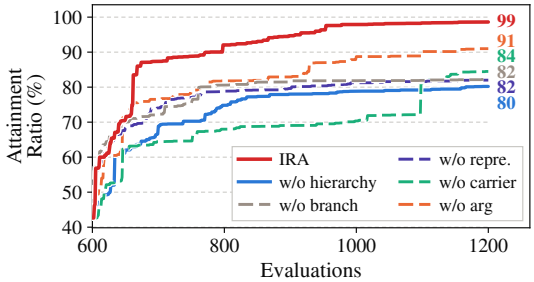


Figure 7: **Analysis of hierarchical revisions in IRA.** Curves show the variation in the mean attainment ratio of the incumbent across six representative *QM9-sample* tasks.

6.4 Dependence on the search learner

To assess whether the descriptors designed by *Pangaea* are overly specialized to the learner used during search, we hold the selected descriptor pipeline for each prediction target fixed and evaluate its descriptors with alternative learners. Figure 8 shows that the designed descriptors are not strongly tied to the learner used during search, outperforming the deployment learner's *Topology & Global* baseline in 51 of 72 cross-learner evaluations. Specifically, descriptors selected with SVR or Ridge remain beneficial in 41 of 48 cross-learner evaluations, whereas those selected with XGBoost do so in 10 of 24. The latter failures are concentrated on the four thermodynamic-energy tasks, where XGBoost itself achieves only modest search gains (at most 4.3%). This suggests that more effective descriptors identified during search tend to capture more task-relevant spatial information and therefore exhibit weaker dependence on the search learner.

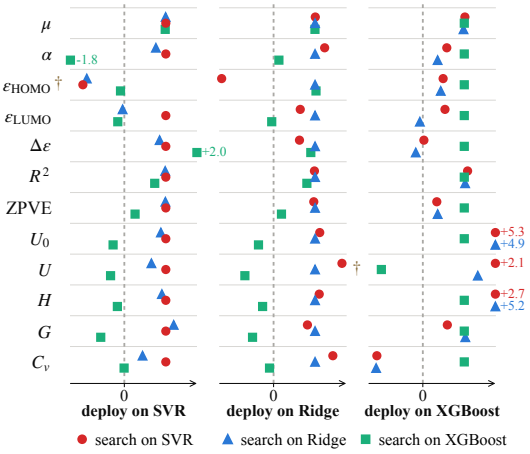


Figure 8: **Dependence of *Pangaea*-designed descriptors on the search learner.** Values are normalized to the *Topology & Global* baseline (0) and the pipeline searched using the deployment learner (1). † denotes underperformance of the latter, in which case the normalized value is −1.

6.5 Analysis of cross-task reuse

To analyze the cross-task reusability of task-specific descriptors, we transfer each pipeline designed for one property unchanged to the other tasks using SVR as the learner, without additional search. As shown in Figure 9, cross-task reuse exhibits structure. Descriptors designed for thermodynamic targets U_0, U, H, and G retain nearly the full own-task gains when transferred within the group, whereas pipelines selected outside this group do not improve these targets. Cross-task transfer is not always symmetric, as seen for c_v and ZPVE, where the c_v pipeline nearly matches the task-specific gain on ZPVE while the reverse transfer is weaker, possibly reflecting related but non-identical information requirements. These results show that task-specific search can yield descriptors reusable across related properties.

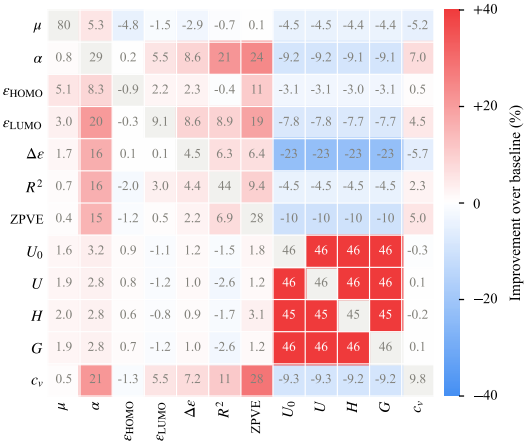


Figure 9: **Cross-task reuse of task-specific descriptors in *Pangaea*.** Each cell reports the improvement (%) over the column task's *Topology & Global* baseline using the descriptor pipeline designed for the row task.

7 Conclusion

In this paper, we abstract the construction of molecular 3D descriptors into a unified pipeline comprising the Formation, Extraction, and Compression stages, recasting task-specific descriptor design as identifying suitable combinations of stage-wise operator choices and parameters. Building on this, we introduce *Pangaea*, which organizes compatible combinations into executable pipelines through a Unified Construction Space and uses an Incumbent Revision Agent to identify task-suited combinations based on evaluation history, design-space semantics, and chemical priors. We validate *Pangaea* in two data-limited settings. On *QM9-sample*, it ranks first among descriptor-based methods for all ML learners, with its best learner surpassing DL methods on 11/12 tasks. On the Fe Mössbauer data, it further improves prediction beyond expert-crafted coordination descriptors. Furthermore, *Pangaea* can integrate new chemical sources and operators with existing components, enabling more combinations for descriptor design across diverse molecular prediction tasks.

AI USE STATEMENT

Generative AI tools assisted with software implementation and debugging and with editing portions of the manuscript for structure and readability. The authors reviewed the AI-assisted code and text incorporated into the work. The authors take responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the aid of generative AI.

REFERENCES

Carlo Adamo and Vincenzo Barone. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.*, 110(13):6158–6170, 1999.

M. Stuart Armstrong, Garrett M. Morris, Paul W. Finn, et al. ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics. *J. Comput. Aided Mol. Des.*, 24(9): 789–801, 2010.

Gustavo A. Arteca. Molecular Shape Descriptors. In Kenny B. Lipkowitz and Donald B. Boyd (eds.), *Reviews in Computational Chemistry*, volume 9, pp. 191–253. John Wiley & Sons, 1996.

Kenneth Atz, David F. Nippa, Alex T. Müller, et al. Geometric deep learning-guided Suzuki reaction conditions assessment for applications in medicinal chemistry. *RSC Med. Chem.*, 15(7):2310–2321, 2024.

Mahendra Awale, Xian Jin, and Jean-Louis Reymond. Stereoselective Virtual Screening of the ZINC Database Using Atom Pair 3D-Fingerprints. *J. Cheminform.*, 7(1):3, 2015.

Malkeet Singh Bahia, Omer Kaspi, Meir Touitou, et al. A Comparison Between 2D and 3D Descriptors in QSAR Modeling Based on Bio-Active Conformations. *Mol. Inform.*, 42(4):e2200186, 2023.

Alexandru T. Balaban. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.*, 89(5):399–404, 1982.

Trent Barnard, Steven Tseng, James P. Darby, et al. Leveraging genetic algorithms to maximise the predictive capabilities of the SOAP descriptor. *Mol. Syst. Des. Eng.*, 8(3):300–315, 2023.

Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Phys. Rev. B*, 87(18):184115, 2013.

Axel D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.*, 98(7):5648–5652, 1993.

Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.*, 134(7):074106, 2011.

Jörg Behler and Michele Parrinello. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.*, 98(14):146401, 2007.

James Bergstra, Rémi Bardenet, Yoshua Bengio, et al. Algorithms for hyper-parameter optimization. In *NeurIPS*, 2011.

Steven H. Bertz. The First General Index of Molecular Complexity. *J. Am. Chem. Soc.*, 103(12): 3599–3601, 1981.

G. Richard Bickerton, Gaia V. Paolini, Jérémy Besnard, et al. Quantifying the Chemical Beauty of Drugs. *Nat. Chem.*, 4:90–98, 2012.

J. Stephen Binkley, John A. Pople, and Warren J. Hehre. Self-consistent molecular orbital methods. 21. Small split-valence basis sets for first-row elements. *J. Am. Chem. Soc.*, 102(3):939–947, 1980.

A. Bondi. van der Waals Volumes and Radii. *J. Phys. Chem.*, 68(3):441–451, 1964.

Johnathan W. Campbell and Konstantinos D. Vogiatzis. Augmenting MACCS Keys with Persistent Homology Fingerprints for Protein–Ligand Binding Classification. *J. Chem. Inf. Model.*, 65(15): 8113–8126, 2025.

Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *KDD*, 2016.

Matteo Cobelli and Stefano Sanvito. Agentic Design of Compositional Descriptors via Autoresearch for Materials Science Applications. arXiv, 2026.

Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Appl. Comput. Harmon. Anal.*, 21(1): 5–30, 2006.

Viviana Consonni, Roberto Todeschini, and Manuela Pavan. Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors. *J. Chem. Inf. Comput. Sci.*, 42(3):682–692, 2002.

Richard D. Cramer, David E. Patterson, and Jeffrey D. Bunce. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.*, 110 (18):5959–5967, 1988.

G. Cruciani, P. Crivori, P.-A. Carrupt, et al. Molecular fields in quantitative structure–permeation relationships: the VolSurf approach. *J. Mol. Struct.: THEOCHEM*, 503(1-2):17–30, 2000.

Sandip De, Albert P. Bartók, Gábor Csányi, et al. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.*, 18(20):13754–13769, 2016.

Jianyuan Deng, Zhibo Yang, Hehe Wang, et al. A systematic study of key elements underlying molecular property prediction. *Nat. Commun.*, 14(1):6395, 2023.

Harris Drucker, Christopher J. C. Burges, Linda Kaufman, et al. Support vector regression machines. In *NeurIPS*, 1996.

Simon S. Du, Wei Hu, Sham M. Kakade, et al. Few-shot learning via learning the representation, provably. In *ICLR*, 2021.

Joseph L. Durant, Burton A. Leland, Douglas R. Henry, et al. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.*, 42(6):1273–1280, 2002.

Felix A. Faber, Luke Hutchison, Bing Huang, et al. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput.*, 13(11):5255–5264, 2017.

Nicholas C. Firth, Nathan Brown, and Julian Blagg. Plane of Best Fit: A Novel Method to Characterize the Three-Dimensionality of Molecules. *J. Chem. Inf. Model.*, 52(10):2516–2525, 2012.

Michael J. Frisch, John A. Pople, and J. Stephen Binkley. Self-consistent molecular orbital methods 25. Supplementary functions for Gaussian basis sets. *J. Chem. Phys.*, 80(7):3265–3269, 1984.

Yu-Chen Gao, Xiang Chen, Yu-Hang Yuan, et al. Accelerating battery innovation: AI-powered molecular discovery. *Chem. Soc. Rev.*, 54(21):9630–9684, 2025.

M. Gastegger, L. Schwiedrzik, M. Bittermann, et al. wACSF—Weighted atom-centered symmetry functions as descriptors in machine learning potentials. *J. Chem. Phys.*, 148(24):241709, 2018.

Johann Gasteiger and Mario Marsili. Iterative Partial Equalization of Orbital Electronegativity—A Rapid Access to Atomic Charges. *Tetrahedron*, 36(22):3219–3228, 1980.

P. J. Goodford. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.*, 28(7):849–857, 1985.

Stefan Grimme, Jens Antony, Stephan Ehrlich, et al. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.*, 132(15):154104, 2010.

Stefan Grimme, Stephan Ehrlich, and Lars Goerigk. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.*, 32(7):1456–1465, 2011.

Jose Carlos Gómez-Tamayo, Joris Tavernier, Roy Aerts, et al. MolAgent: Biomolecular Property Estimation in the Agentic Era. *J. Chem. Inf. Model.*, 65(20):10808–10818, 2025.

Lowell H. Hall and Lemont B. Kier. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure–Property Modeling. In Kenny B. Lipkowitz and Donald B. Boyd (eds.), *Reviews in Computational Chemistry*, volume 2, pp. 367–422. Wiley-VCH, 1991.

Nikolaus Hansen and Andreas Ostermeier. Completely Derandomized Self-Adaptation in Evolution Strategies. *Evol. Comput.*, 9(2):159–195, 2001.

Ruolin He, Shaohua Gu, Jiazheng Xu, et al. SIDERITE: Unveiling hidden siderophore diversity in the chemical space through digital exploration. *iMeta*, 3(2):e192, 2024.

Markus C. Hemmer, Valentin Steinhauer, and Johann Gasteiger. Deriving the 3D structure of organic molecules from their infrared spectra. *Vib. Spectrosc.*, 19(1):151–164, 1999.

Alberto Hernandez, Adarsh Balasubramanian, Fenglin Yuan, et al. Fast, accurate, and transferable many-body interatomic potentials by symbolic regression. *npj Comput. Mater.*, 5(1):112, 2019.

Lauri Himanen, Marc O.J. Jäger, Eiaki V. Morooka, et al. DScribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.*, 247:106949, 2020.

Arthur E. Hoerl and Robert W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67, 1970.

Haoyan Huo and Matthias Rupp. Unified representation of molecules and crystals for machine learning. *Mach. Learn.: Sci. Technol.*, 3(4):045017, 2022.

Dejun Jiang, Zhenxing Wu, Chang-Yu Hsieh, et al. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *J. Cheminform.*, 13(1):12, 2021.

Alan R. Katritzky, Lan Mu, Victor S. Lobanov, et al. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics. *J. Phys. Chem.*, 100(24):10400–10407, 1996.

James Kennedy and Russell Eberhart. Particle swarm optimization. In *ICNN*, 1995.

Lemont B. Kier and Lowell H. Hall. An Electrotopological-State Index for Atoms in Molecules. *Pharm. Res.*, 7(8):801–807, 1990.

Gerhard Klebe and Ute Abraham. Comparative Molecular Similarity Index Analysis (CoMSIA) to Study Hydrogen-Bonding Properties and to Score Combinatorial Libraries. *J. Comput. Aided Mol. Des.*, 13(1):1–10, 1999.

Gerhard Klebe, Ute Abraham, and Thomas Mietzner. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity. *J. Med. Chem.*, 37(24):4130–4146, 1994.

Ernő Kuzmann, Zoltán Homonnay, Zoltán Klencsár, et al. ^{57}Fe Mössbauer Spectroscopy as a Tool for Study of Spin States and Magnetic Interactions in Inorganic Chemistry. *Molecules*, 26(4): 1062, 2021.

Paul Labute. A Widely Applicable Set of Descriptors. *J. Mol. Graphics Modell.*, 18(4–5):464–477, 2000.

Greg Landrum et al. RDKit: Open-source cheminformatics. `https://doi.org/10.5281/zenodo.19250388`, 2026. Version 2026.03.1.

Marcel F. Langer, Alex Goeßmann, and Matthias Rupp. Representations of Molecules and Materials for Interpolation of Quantum-Mechanical Simulations via Machine Learning. *npj Comput. Mater.*, 8(1):41, 2022.

Jeffrey Larson, Matt Menickelly, and Stefan M. Wild. Derivative-free optimization methods. *Acta Numer.*, 28:287–404, 2019.

Byungkook Lee and Frederic M. Richards. The Interpretation of Protein Structures: Estimation of Static Accessibility. *J. Mol. Biol.*, 55(3):379–400, 1971.

Chengteh Lee, Weitao Yang, and Robert G. Parr. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, 37(2):785–789, 1988.

Tian Lu. A Comprehensive Electron Wavefunction Analysis Toolbox for Chemists, Multiwfn. *J. Chem. Phys.*, 161(8):082503, 2024.

Tian Lu and Feiwu Chen. Atomic Dipole Moment Corrected Hirshfeld Population Method. *J. Theor. Comput. Chem.*, 11(1):163–183, 2012a.

Tian Lu and Feiwu Chen. Multiwfn: A Multifunctional Wavefunction Analyzer. *J. Comput. Chem.*, 33(5):580–592, 2012b.

Łukasz Mentel. mendeleev – A Python resource for properties of chemical elements, ions and isotopes. `https://github.com/lmmentel/mendeleev/releases/tag/v1.2.0`, 2026. Version 1.2.0.

Grégoire Montavon, Katja Hansen, Siamac Fazli, et al. Learning Invariant Representations of Molecules for Atomization Energy Prediction. In *NeurIPS*, 2012.

Felix Musil, Andrea Grisafi, Albert P. Bartók, et al. Physics-Inspired Structural Representations for Molecules and Materials. *Chem. Rev.*, 121(16):9759–9815, 2021.

Jingjie Ning, Xiaochuan Li, Ji Zeng, et al. Closed-loop Auto Research for Molecular Property Prediction: Discovering and Certifying Generalizable Improvements. arXiv, 2026.

Artem R. Oganov and Mario Valle. How to quantify energy landscapes of solids. *J. Chem. Phys.*, 130(10):104504, 2009.

Runhai Ouyang, Stefano Curtarolo, Emre Ahmetcik, et al. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Phys. Rev. Mater.*, 2(8):083802, 2018.

Manuel Pastor, Gabriele Cruciani, Iain McLay, et al. GRid-INdependent Descriptors (GRIND): A Novel Class of Alignment-Independent Three-Dimensional Molecular Descriptors. *J. Med. Chem.*, 43(17):3233–3243, 2000.

R. S. Pearlman and K. M. Smith. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.*, 39(1):28–35, 1999.

Max Pinheiro, Fuchun Ge, Nicolas Ferré, et al. Choosing the right molecular machine learning potential. *Chem. Sci.*, 12(43):14396–14413, 2021.

Pekka Pyykkö and Michiko Atsumi. Molecular Single-Bond Covalent Radii for Elements 1–118. *Chem. Eur. J.*, 15(1):186–197, 2009.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, et al. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data*, 1:140022, 2014.

Jiahua Rao, Dahao Xu, Wentao Wei, et al. Quadruple Attention in Many-body Systems for Accurate Molecular Property Predictions. In *ICML*, 2025.

Luis Miguel Rios and Nikolaos V. Sahinidis. Derivative-free optimization: a review of algorithms and comparison of software implementations. *J. Glob. Optim.*, 56(3):1247–1293, 2013.

David Rogers and Mathew Hahn. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.*, 50(5): 742–754, 2010.

Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, et al. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.*, 108(5):058301, 2012.

Soumitra Samanta, Bob Luigjes, Stuart Lyon, et al. Novel corrosion protective coating formulations suggested by machine learning of historic test data. *npj Mater. Degrad.*, 10(1):63, 2026.

Wolfgang H. B. Sauer and Matthias K. Schwarz. Molecular Shape Diversity of Combinatorial Libraries: A Prerequisite for Broad Bioactivity. *J. Chem. Inf. Comput. Sci.*, 43(3):987–1003, 2003.

Kristof T. Schütt, Pieter-Jan Kindermans, Huziel E. Sauceda, et al. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *NeurIPS*, 2017.

Jan H. Schuur, Paul Selzer, and Johann Gasteiger. The Coding of the Three-Dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure-Spectra Correlations and Studies of Biological Activity. *J. Chem. Inf. Comput. Sci.*, 36(2):334–344, 1996.

Mark A. Spackman and Patrick G. Byrom. A Novel Definition of a Molecule in a Crystal. *Chem. Phys. Lett.*, 267(3–4):215–220, 1997.

Afnan Sultan, Max Rausch-Dupont, Shahrukh Khan, et al. Transformers for molecular property prediction: domain adaptation efficiently improves performance. *J. Cheminform.*, 18(1):107, 2026.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Roberto Todeschini and Viviana Consonni. *Molecular Descriptors for Chemoinformatics*. Wiley-VCH, Weinheim, 2 edition, 2009. ISBN 978-3-527-31852-0.

Roberto Todeschini, Marina Lasagni, and Emilio Marengo. New Molecular Descriptors for 2D and 3D Structures. Theory. *J. Chemom.*, 8(4):263–272, 1994.

Nilesh Tripuraneni, Michael I. Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. In *NeurIPS*, 2020.

Markus Wagener, Jens Sadowski, and Johann Gasteiger. Autocorrelation of Molecular Surface Properties for Modeling Corticosteroid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc.*, 117(29):7769–7775, 1995.

Liang Wang, Shaozhen Liu, Yu Rong, et al. MolSpectra: Pre-training 3D Molecular Representation with Multi-modal Energy Spectra. In *ICLR*, 2025.

Yusong Wang, Tong Wang, Shaoning Li, et al. Enhancing geometric representations for molecules with equivariant vector-scalar interactive message passing. *Nat. Commun.*, 15(1):313, 2024.

Florian Weigend and Reinhart Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.*, 7(18):3297–3305, 2005.

Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, et al. Large language models for scientific discovery in molecular property prediction. *Nat. Mach. Intell.*, 7(3):437–447, 2025.

APPENDIX CONTENTS

A Decomposition of existing 3D molecular descriptors 16
B Details of the Incumbent Revision Agent 18
B.1 Evaluation history retrieval 18
B.2 Hierarchical revisions 18
C Dataset construction 18
D Baseline details 19
D.1 ML learner configurations 19
D.2 *Topology & Global* descriptors for *QM9-sample* 19
D.3 *Coordination & Global* descriptors for *MB* 21
D.4 Hand-crafted 3D descriptor baselines 22
D.5 End-to-end DL baselines 23
E Implementation details 23
E.1 Source library instantiation 23
E.2 Address instantiation 23
E.3 Closed-loop search protocol 25
E.4 Black-box optimizers as search strategies 25
E.5 Computational environment 25
F Additional results for the analysis of IRA 25
F.1 Search-strategy comparison 25
F.2 Analysis of search guidance in IRA 26
F.3 Analysis of hierarchical revisions in IRA 27
G Pipelines designed by *Pangaea* 27
G.1 Designed pipelines for *QM9-sample* 27
G.2 Visual analysis of the pipelines designed by *Pangaea* for *MB* 32
H Details of the Unified Construction Space 33
H.1 Shared source library 33
H.2 Branch selection and address layout 34
H.3 Common pipeline notation 34
H.4 Point branch 35
H.5 Field branch 39
H.6 Expansion branch 42
H.7 Statistics branch 45

A DECOMPOSITION OF EXISTING 3D MOLECULAR DESCRIPTORS

We survey twenty representative classes of existing 3D descriptors from established descriptor libraries and the primary literature, focusing on molecular descriptors and including a crystal fingerprint as an extension. Table S1 shows the decomposition of these descriptors, organized into four groups by construction pattern.

Table S1: Decomposition details of existing 3D molecular descriptors.

Descriptor	Formation	Extraction	Compression
Point			
Coulomb matrix (CM) (Rupp et al., 2012; Montavon et al., 2012)	Nuclear charges Z_i at atomic positions $\mathbf{R}_i$ define an attributed point cloud.	Atom-pair matrix with diagonal entries $M_{ii} = 0.5Z_i^{2.4}$ and off-diagonal entries $M_{ij} = Z_iZ_j/\|\mathbf{R}_i - \mathbf{R}_j\|_2$	Sort eigenvalues by decreasing absolute value and zero-pad to a fixed length; or jointly permute rows and columns by decreasing row norm, pad the matrix with zeros to a fixed size, and vectorize
Solvent-accessible surface area (ASA) / van der Waals volume[1] (Lee & Richards, 1971; Bondi, 1964)	Atomic positions and van der Waals radii define a union of atom-centered balls. For ASA, each radius is increased by the probe radius.	Identify unoccluded spherical boundary patches and compute their areas (ASA); sample parallel planar sections and compute their union areas with integration weights (volume)	Sum patch areas (ASA) or integrate cross-sectional areas by weighted summation (volume); return a scalar for each quantity
Gravitational index (Katritzky et al., 1996)	Atomic masses m_i and coordinates define a point cloud under the prescribed hydrogen convention. Bond connectivity is retained for bonded-pair variants.	Enumerate unordered pairs $i < j$, either all pairs or only bonded pairs; compute their Euclidean distances r_{ij} and mass products m_im_j	Sum m_im_j/r_{ij}^2 over each selected pair set; return the sum, its square root, or its cube root, concatenating the chosen variants in fixed order
Field			
CoMFA[2] (Cramer et al., 1988)	Lennard–Jones steric and Coulombic electrostatic interactions between an aligned molecule and a probe atom define two energy fields.	Sample both fields on a shared 3D lattice, applying energy cutoffs and electrostatic treatment in strongly repulsive regions	Vectorize and concatenate field samples in fixed order
CoMSIA[2] (Klebe et al., 1994; Klebe & Abraham, 1999)	Probe similarities for an aligned molecule define Gaussian-attenuated fields.	Sample steric, electrostatic, hydrophobic, and hydrogen-bond donor and acceptor fields on a shared regular 3D lattice	Vectorize and concatenate field samples in fixed order
GRID[3] / VolSurf (Goodford, 1985; Cruciani et al., 2000)	Van der Waals, electrostatic, and applicable hydrogen-bond interactions between a molecular conformation and chemical probes define the GRID interaction-energy fields.	Sample each probe-specific field on a 3D grid, retaining grid coordinates, interaction energies, and probe identities	Compute VolSurf size and shape indices, hydrophilic and hydrophobic region volumes at prescribed energy levels, and their ratios and spatial-distribution measures; concatenate into a fixed-length vector
GRIND (Pastor et al., 2000)	Interactions between a molecular conformation and chemical probes define probe-specific molecular interaction fields around the molecule.	Select representative grid nodes within each field by balancing favorable interaction energies and spatial separation; form within- and cross-field node pairs with Euclidean distances, energy products, and probe-pair labels	Retain the maximum node-energy product in each prescribed distance bin for each probe pair; concatenate the auto- and cross-correlograms
Expansion			
PMI / NPR / R_g / asphericity / PBF (Sauer & Schwarz, 2003; Arteca, 1996; Firth et al., 2012)	Mass or unit weights at atomic positions define the point cloud for moment-based descriptors. PBF uses an unweighted point cloud of the heavy atoms.	Compute centered inertia tensors for PMI/NPR and normalized gyration tensors for radius of gyration and asphericity, using centers consistent with the weights; for PBF, fit a least-squares plane to heavy atoms and compute their unsigned perpendicular distances	Return sorted principal moments and ratios to the largest moment; compute radius of gyration and normalized asphericity from gyration-tensor eigenvalues; average the distances for PBF
WHIM (Todeschini et al., 1994)	Unit and selected atomic-property weights at atomic positions define an attributed point cloud.	Center the coordinates and perform PCA of the weighted covariance matrix for each weighting scheme; retain ordered eigenvalues and scores along the principal axes	Compute directional and global indices of size, shape, symmetry, and atom distribution from the eigenvalues and atomic scores; concatenate across weighting schemes
3D-MoRSE (Schuur et al., 1996)	Atomic-property weights w_i at atomic coordinates define a weighted point cloud.	Compute Euclidean distances r_{ij} and weight products w_iw_j for all atom pairs $i < j$; specify a fixed set of scattering parameters s	For each s, sum $\sin(u)/u$ over $i < j$ with weights w_iw_j, where $u = sr_{ij}$; use the kernel limit 1 at $u = 0$ and concatenate across s and weighting schemes
SOAP[3] (Bartók et al., 2013; De et al., 2016)	Element labels and relative atomic coordinates define element-resolved densities around each atom by summing Gaussian contributions from the center and its neighbors, using a smooth radial cutoff.	Expand each density in radial basis functions and spherical harmonics using fixed element channels and radial and angular truncation orders; retain the coefficients for each center	Contract coefficient products over magnetic indices to form rotationally invariant power spectra for each element pair; concatenate and normalize each local vector, then apply a fixed permutation-invariant pooling over centers

Table S1 (continued)

Descriptor	Formation	Extraction	Compression
GETAWAY (H and R, including HATS) (Consonni et al., 2002)	Unit or carbon-scaled atomic-property weights at atomic positions define a point cloud including hydrogens. The cloud is centered at its geometric centroid and retains bond connectivity.	Compute the molecular influence matrix H from centered coordinates using a pseudoinverse; form R from geometric means of paired leverages divided by Euclidean distances, with zero diagonal; compute graph shortest-path distances	Compute leverage statistics, information indices, and global R-matrix indices; form property-weighted H autocorrelations using positive h_{ij}, HATS autocorrelations using $w_i h_{ii}$, and weighted R sums and maxima by graph-distance lag; concatenate selected indices, totals, and fixed-lag profiles across weighting schemes
Statistics			
ElectroShape (Armstrong et al., 2010)	Atomic partial charges, scaled by a fixed conversion factor and attached to atomic positions, define the attributed point cloud.	Use the 4D centroid and two successively farthest atoms as the first three reference points; form two more using a shared scaled spatial cross-product offset from the centroid and the maximum/minimum charge coordinates; compute 4D Euclidean atom-to-reference distances	Compute the mean, standard deviation, and real cube root of the third central moment of each distance distribution; concatenate into a 15-component vector
RDF code (Hemmer et al., 1999)	Weights from the selected atomic weighting schemes, attached to atomic positions, define the attributed point cloud.	Enumerate all unordered pairs of distinct atoms; compute their Euclidean distances r_{ij} and weight products $w_i w_j$	At each fixed radius r, sum $w_i w_j \exp[-B(r - r_{ij})^2]$ over $i < j$ with fixed $B > 0$; concatenate across radii and schemes
Autocorrelation of molecular surface properties (AMSP) (Wagener et al., 1995)	Selected surface-property values on points randomly sampled from the molecular van der Waals surface at a prescribed density define the attributed point cloud.	Enumerate unordered pairs of distinct surface points; compute their Euclidean distances and the products of their values for each selected surface property	For each prescribed distance interval, average the property products over the point pairs in that interval; concatenate over a fixed set of intervals and surface properties
Valle–Oganov crystal fingerprint (partial RDF)[4] (Oganov & Valle, 2009)	Element-labeled atomic coordinates and lattice vectors define a periodically extended crystal point cloud for a fixed radial cutoff, retaining the reference-cell volume and species counts.	Enumerate pairs between reference-cell atoms and atoms in the periodically extended crystal within the cutoff; exclude zero-distance self-pairs, compute Euclidean separations, and resolve them by unordered element pair	Sum Gaussian peaks with spherical shell normalization; divide by central species counts per cell and neighbor number densities; subtract one and concatenate channels on a fixed radial grid
3D atom-pair fingerprints (3DAPfp/3DXfp) (Awale et al., 2015)	Heavy-atom coordinates define the point cloud. For 3DXfp, its points also carry potentially overlapping hydrophobic (Hyb), hydrogen-bond donor (HBD), hydrogen-bond acceptor (HBA), and planar (sp^2) categories.	Compute Euclidean distances between distinct heavy atoms, using all pairs for 3DAPfp or the four same-category pair types and the HBA–HBD cross-pair for 3DXfp; sample a Gaussian centered at each distance, with width $0.18 d_{ij}$, at 16 fixed geometrically spaced distance values	Sum sampled contributions within each channel; apply the prescribed global or category-specific atom-count normalization, multiply by 100, round to integers, and concatenate channels
ACSF / wACSF[3] (Behler & Parrinello, 2007; Behler, 2011; Gastegger et al., 2018)	Element labels and coordinates of neighbors within a fixed cutoff define a local point cloud around each atom, with the central atom's element label retained.	For each center, evaluate fixed sets of radial and angular symmetry functions by summing two-body and three-body terms with smooth cutoff factors; separate neighbor-element and neighbor-element-pair channels for ACSF, or weight radial and angular terms by Z_j and $Z_j Z_k$ for wACSF	Concatenate symmetry-function values into a vector for each center; apply fixed permutation-invariant pooling within each central-element group, then concatenate groups in a fixed element order, using zero blocks for absent elements
Local MBTR (LMBTR)[3,5] (Huo & Rupp, 2022)	Element labels and atomic coordinates define a local point cloud for each atomic center. The center is encoded as a virtual species X, with its original element retained separately.	Enumerate distinct-site pairs and triples containing X; compute pair inverse distances and triplet angle cosines, resolving channels by element tuple and distinguishing X at the vertex from X at an endpoint	Sum weighted Gaussian contributions on fixed grids per channel and center; apply permutation-invariant pooling by original central element; concatenate channels and element groups in fixed order, zero-filling absent groups
Histograms of distances, angles, and dihedrals (HD/HDA/HDAD) (Faber et al., 2017)	Element labels at atomic positions define an attributed point cloud that includes hydrogen atoms.	Compute all unordered atom-pair distances; for HDA/HDAD, compute angles at each atom from pairs of its three nearest neighbors; for HDAD, use each of its four nearest neighbors in turn to define a center–neighbor axis and compute dihedrals between planes formed with pairs of the remaining neighbors; label channels by element combinations, distinguishing the center element for angular terms	Linearly distribute each value over the two nearest predefined bin centers in its element channel; sum contributions per bin across the molecule and concatenate histograms in a fixed channel and bin order

[1] For ASA and van der Waals volume, the table describes geometric constructions of the corresponding quantities; the numerical discretization need not reproduce the original implementations.

[2] Conventional CoMFA and CoMSIA fit partial least squares to the flattened lattice vectors, with cross-validation in CoMFA. When descriptor and predictor are separated, as in our cross-learner experiments, the pipeline output is the flattened descriptor vector, and the supervised projection belongs to the downstream model.

[3] A construction adapted to a fixed-length molecular representation. SOAP, ACSF/wACSF, and LMBTR each natively return one fixed-width vector per center; the adaptation applies permutation-invariant pooling over centers within compression.

[4] This row describes the native global fingerprint of periodic crystals, included as an extension of the three-stage decomposition to periodic structures.

[5] LMBTR follows the terminology used in the DScribe library (Himanen et al., 2020), with the local construction specified in its official LMBTR documentation: it is the local specialization of MBTR obtained by requiring every retained tuple to contain a designated center, not an independently derived global descriptor.

B Details of the Incumbent Revision Agent

B.1 Evaluation history retrieval

Global tools. There are four types of global tools. *Leaders* lists the highest-scoring distinct pipelines overall and within each branch. *Coverage* summarizes how evaluations are distributed across branches and operator choices, and identifies unexplored parts of the operator-choice space. *Recent* summarizes recent evaluations and their associated proposal outcomes, including evaluations that produce a new highest-scoring pipeline. *Frontier* lists up to twenty pipelines on the Pareto frontier of search score and rarity, where rarity is the mean reciprocal occurrence count of the pipeline's active operator choices and parameter bins among distinct evaluated pipelines in the same branch.

Targeted tools. IRA is provided with five targeted tools for retrieving additional information as needed and may invoke up to two of them in each round. *Record* returns an evaluated pipeline's address, decoded specification, and search score. *Slice* summarizes pipelines matching specified operator choices and discrete operator parameters, reporting their distinct count, score quantiles, diversity across the remaining discrete dimensions, continuous-parameter coverage, and best-performing pipelines. *Branch* provides a detailed summary of a specified branch, including evaluation counts and coverage of operator choices and parameters, along with its highest-scoring distinct pipelines. *Neighbor* returns distinct pipelines closest to a specified reference pipeline, ranked first by the number of differences in branch assignment, operator choices, and discrete operator parameters, and then by differences in continuous-parameter regions, together with their score differences from the reference. *Revise* retrieves earlier proposals relevant to a specified search context, including their revision rationales, recorded outcomes, and conditions for reconsideration.

B.2 Hierarchical revisions

The four revision levels differ in their primary targets and the settings they preserve. Table S2 summarizes the revision scope at each level.

Table S2: Scope of revisions at each level.

Revision level	Revision scope
Branch	Switches to another branch and configures a complete, compatible pipeline within it.
Representation	Retains the branch and focuses on representation-level method selection, primarily through the Extraction and Compression operators. Includes a change to at least one of these operators and a change to a continuous-valued parameter, while allowing other unfixed operator choices and parameters to adjust jointly.[1]
Carrier	Retains the incumbent's branch and the operator choices established at the representation level, focusing primarily on carrier-related changes. Includes a change to the Formation operator or one of its parameters, together with a continuous-parameter change. When appropriate, a change to another unfixed discrete parameter can replace the Formation change.
Argument	Modifies at least one continuous parameter, while keeping the branch, operators and discrete parameters fixed.

[1] In the *field branch*, although kernel type is technically a parameter of the Formation operator, it is treated as a representation-level choice because of its defining role in the representation used by that branch.

C Dataset construction

***QM9-sample*.** To study molecular-property learning with limited data, we construct a fixed 1498-molecule subset from the 130,831-molecule QM9 pool. All molecules with 1–4 heavy atoms are retained, while the 5–9 heavy-atom groups are capped at $\{100, 150, 250, 350, 600\}$ molecules, respectively. Within each capped group, molecules are sampled in a round-robin manner across HOMO–gap quantile cells and Murcko scaffolds, with preference given to rare scaffolds to prevent any single scaffold or property region from dominating the subset. The subset is then split into a fixed train/validation/test partition of 899/300/299 molecules, respectively. Specifically, we define 24 strata from six heavy-atom-count groups (pooling molecules with 1–4 heavy atoms and treating those with 5–9 separately) and two quantile bins each for $\varepsilon_{\mathrm{HOMO}}$ and $\Delta\varepsilon$. Within each stratum, we match similar molecules into quintets wherever possible, yielding 291 quintets and 43 unmatched

molecules because the remainders are not pooled across strata. To form the split, we allocate three molecules from each quintet to the training set and one each to the validation and test sets. We then assign the unmatched molecules individually to sets with remaining capacity, favoring assignments that balance the distributions of heavy-atom count and target properties across the sets. Each candidate's search score is the negative of the mean MAE across the five validation folds on the pooled training and validation sets (1199 molecules). After selection, the pipeline is fixed and the learner is refitted using the baseline descriptors augmented with the selected pipeline's descriptors on the original training set (899 molecules) and evaluated on the held-out test set (299 molecules).

MB. *MB* comprises experimental Fe Mössbauer spectroscopy measurements curated from the literature, covering 617 Fe sites from 495 molecules. Each site has four targets: the isomer shift δ (mm/s), which reflects the s-electron density at the iron nucleus; the magnitude of the quadrupole splitting ΔE_Q (mm/s), determined by the electric-field gradient at the nucleus; the formal oxidation state of iron, ranging from Fe(0) to Fe(IV); and the iron spin state, categorized as 0, $\frac{1}{2}$, 1, $\frac{3}{2}$, 2, or $\frac{5}{2}$. The first two are regression targets and the last two are classification targets. The dataset is split into a fixed train/validation/test partition of 205/206/206 sites, respectively, with all Fe sites from the same molecule assigned to the same subset. For each candidate, the learner is fitted on the training set, and the search score is the negative validation MAE for the two regression targets or the negative validation log-loss for the two classification targets. After selection, the pipeline is fixed and the learner is refitted using the baseline descriptors augmented with the selected pipeline's descriptors on the original training set and evaluated on the held-out test set.

D BASELINE DETAILS

D.1 ML LEARNER CONFIGURATIONS

In this paper, we use three ML learners: support vector regression (SVR), a penalized linear model (Ridge), and a gradient-boosted tree ensemble (XGBoost). Each configuration is tuned on the validation split, once per task and learner. Table S3 lists the selected configurations.

Table S3: Learner configurations for *QM9-sample* and *MB*.

Learner	Targets	Selected configuration
SVR	μ	linear kernel, $C = 0.55$, $\varepsilon = 0$
	α	degree-2 polynomial kernel, $C = 5.0$, $\varepsilon = 0.06$, scaled γ
	$\varepsilon_{\mathrm{HOMO}}$, $\varepsilon_{\mathrm{LUMO}}$, $\Delta\varepsilon$	radial-basis kernel, $C = 1.0$, $\varepsilon = 0.001$, $\gamma = 0.001$
	R^2	degree-2 polynomial kernel, $C = 90$, $\varepsilon = 0$, automatic γ, tolerance 0.75
	ZPVE	linear kernel, $C = 0.03$, $\varepsilon = 0.0005$
	U_0, U, H, G	linear kernel, $C = 0.55$, $\varepsilon = 0.2$
	c_v	degree-2 polynomial kernel, $C = 1.15$, $\varepsilon = 0.01$, scaled γ
Ridge	*QM9-sample*, all except c_v	$\alpha = 1$, intercept fitted
	c_v	$\alpha = 3$, intercept fitted
XGBoost	*QM9-sample*, all twelve	depth 4, 250 trees, learning rate 0.03, subsample 0.8, column subsample 0.8, L_2 penalty 1
	MB, all four	depth 8, 250 trees, learning rate 0.03, subsample 0.8, column subsample 0.8, L_2 penalty 1

D.2 *Topology & Global* DESCRIPTORS FOR *QM9-sample*

For *QM9-sample* we construct the *Topology & Global* descriptors, which combine standard molecular-graph and RDKit descriptors (Landrum et al., 2026; Todeschini & Consonni, 2009), composition and substructure counts, and pooled electronic-property summaries. Before pipeline search, we select the task-specific baseline descriptor features using the fixed validation split, once per target. For R^2, the Morgan fingerprint block is reduced by PCA before feature selection. Table S4 lists the descriptors for each task.

Elemental composition and simple molecular counts. These descriptors include H/C/N/O/F counts and fractions, together with total-atom, heavy-atom, bond, ring, and rotatable-bond counts.

Table S4: Task-specific *Topology & Global* features for *QM9-sample*; tasks with identical descriptor sets are grouped. A check mark indicates that at least one feature is retained.

Features	μ/α/ZPVE	$\varepsilon_{\mathrm{HOMO}}$	$\varepsilon_{\mathrm{LUMO}}$	$\Delta\varepsilon$	R^2	$U_0/U/H/G$	c_v
Elemental composition and simple molecular counts	✓	✓	✓	✓	✓	✓	✓
Hybridization, aromaticity, and heteroatom content	✓	✓	✓	✓	✓	✓	✓
E-State, EState–VSA, and VSA–EState	✓	✓	✓	✓	✓	✓	✓
Gasteiger charge descriptors	✓	✓	✓	✓		✓	✓
Hall–Kier α	✓	✓	✓	✓			
Chi connectivity indices		✓	✓				✓
Kappa shape indices		✓	✓	✓			✓
BalabanJ and BertzCT	✓	✓	✓	✓		✓	✓
BCUT2D	✓	✓	✓	✓			✓
PEOE/SMR/SlogP–VSA	✓	✓	✓	✓	✓		✓
Physicochemical and drug-likeness descriptors	✓	✓	✓	✓	✓	✓	
Functional-group and ring-substructure descriptors	✓	✓	✓	✓	✓		✓
Morgan fingerprint density and molecular fingerprints		✓		✓	✓		✓
Graph structure, spectra, and autocorrelations					✓		✓
Electronegativity summaries	✓	✓	✓	✓		✓	
ADCH charge summaries	✓	✓	✓	✓		✓	
Molecular-surface ESP extrema summaries	✓	✓	✓	✓		✓	✓
Fragment-surface electron-density extrema summaries	✓	✓	✓	✓		✓	
ESP-basin extrema summaries	✓	✓	✓	✓		✓	✓
Other task-specific features					Composition/graph features		Wavefunction features and products

Hybridization, aromaticity, and heteroatom content. These descriptors include selected atom counts or fractions by hybridization type, together with selected counts or fractions of aromatic atoms, aromatic bonds, and heteroatoms.

E-State, EState–VSA, and VSA–EState. These descriptors include summary statistics and counts by sign for atomic E-State indices (Kier & Hall, 1990), together with RDKit EState–VSA and VSA–EState descriptors (Landrum et al., 2026).

Gasteiger charge descriptors. These descriptors include summaries and counts by sign of Gasteiger charges (Gasteiger & Marsili, 1980) over all atoms and heavy atoms, together with selected charge extrema.

Hall–Kier α. This descriptor is the Hall–Kier α correction used in Kappa shape indices to account for atomic size and hybridization (Hall & Kier, 1991).

Chi connectivity indices. These descriptors include selected Chi connectivity indices of different orders, based on atomic degrees or valence-electron information (Hall & Kier, 1991; Landrum et al., 2026).

Kappa shape indices. These descriptors include selected Kappa indices of orders one to three that characterize molecular shape from graph connectivity (Hall & Kier, 1991).

BalabanJ and BertzCT. These descriptors include the Balaban J index based on graph distances (Balaban, 1982) and the Bertz index of molecular complexity (Bertz, 1981).

BCUT2D. These descriptors include selected BCUT2D indices combining molecular connectivity with atomic mass, Gasteiger charge, and atomic contributions to logP and molar refractivity (Pearlman & Smith, 1999; Landrum et al., 2026).

PEOE/SMR/SlogP–VSA. These descriptors include sums of approximate atomic surface areas within bins defined by Gasteiger charge or atomic contributions to molar refractivity or logP (Labute, 2000; Landrum et al., 2026).

Physicochemical and drug-likeness descriptors. These descriptors include selected molecular-weight measures, logP, molar refractivity, and TPSA (Landrum et al., 2026), together with QED as a measure of drug-likeness (Bickerton et al., 2012).

Functional-group and ring-substructure descriptors. These descriptors include counts of selected functional groups, such as ketones, amides, ethers, and nitriles, together with ring-substructure features (Landrum et al., 2026).

Morgan fingerprint density and molecular fingerprints. These descriptors include Morgan fingerprint densities and selected features from Morgan bit/count (Rogers & Hahn, 2010), MACCS (Durant et al., 2002), atom-pair, topological-torsion, pattern, and path-based fingerprints (Landrum et al., 2026).

Graph structure, spectra, and autocorrelations. These descriptors include features of local graph environments, graph distance statistics, spectral features of graph matrices, and topological autocorrelations of atomic properties (Todeschini & Consonni, 2009).

Electronegativity summaries. These descriptors include selected summary statistics of tabulated atomic electronegativities for each of six scales (Appendix H.1), together with counts of available and missing values.

ADCH charge summaries. These descriptors include the mean, standard deviation, minimum, maximum, and range of atomic ADCH charges (Lu & Chen, 2012a).

Molecular-surface ESP extrema summaries. These descriptors include means and population variances of ESP values at local maxima and minima on the molecular surface, pooled separately and together (Lu & Chen, 2012b).

Fragment-surface electron-density extrema summaries. These descriptors include means and population variances of electron-density values at local maxima and minima on the fragment Hirshfeld surface (Spackman & Byrom, 1997) (Appendix H.1), pooled separately and together (Lu & Chen, 2012b).

ESP-basin extrema summaries. These descriptors include the mean, standard deviation, minimum, maximum, and range of ESP values at basin extrema (Lu & Chen, 2012b).

Other task-specific features. These descriptors include selected composition and graph indicators and products involving rotatable-bond counts for R^2, together with scalar and orbital quantities derived from DFT wavefunctions and selected products of the scalar quantities with graph or elemental summaries for c_v.

D.3 *Coordination & Global* DESCRIPTORS FOR *MB*

MB contains records for individual Fe sites. Accordingly, we construct the *Coordination & Global* baseline from expert-crafted descriptors designed to characterize the local Fe coordination environment, including its geometry, coordinating-atom composition, and Fe–ligand bond lengths. Feature selection follows the same protocol as for *QM9-sample*. Table S5 lists the retained *Coordination & Global* descriptor families for each task.

Continuous shape measure (CShM). This dimensionless measure quantifies the deviation of the Fe coordination polyhedron from the closest match among seventeen ideal coordination geometries, from linear to tricapped trigonal prismatic. A smaller value indicates a less distorted polyhedron.

RMSD from the reference polyhedron. This is the root-mean-square deviation, in angstroms, between the coordinating-atom positions and the same seventeen ideal geometries.

Coordination number. These features are the σ, π, and total coordination numbers of the Fe site, where the π term counts ligands that coordinate through their π system.

Coordinating-atom counts. These features describe the composition of the first coordination shell through individual counts of N, C, O, P, S, H, Si, Te, Se, As, and B donors and pooled counts of metal and halogen donors.

Fe–ligand bond lengths and counts. These features comprise the minimum, maximum, mean, and standard deviation of the Fe–donor distances, together with the number of coordination bonds.

Table S5: Task-specific *Coordination & Global* feature families for the four *MB* tasks. A check mark indicates that at least one feature is retained.

Feature family	δ	ΔE_Q	Iron ox. state	Iron spin state
Continuous shape measure (CShM)	✓		✓	
RMSD from the reference polyhedron		✓		
Coordination number	✓		✓	✓
Coordinating-atom counts	✓	✓	✓	✓
Fe–ligand bond lengths and counts		✓	✓	✓
Bond-length deviation from reference	✓	✓	✓	✓
Coordination-shell sphericity and volume			✓	✓
Ligand electronegativity summaries	✓	✓	✓	✓
Spin state and oxidation state	✓	✓		
Measured Mössbauer spectral parameters			✓	✓
Wavefunction-derived electronic summaries				✓

Bond-length deviation from reference. These features comprise the minimum, maximum, mean, and standard deviation of the percentage deviation of each Fe–donor distance from the sum of the Pyykkö single-bond covalent radii (Pyykkö & Atsumi, 2009); positive values indicate elongated bonds and negative values indicate compressed bonds.

Coordination-shell sphericity and volume. These features comprise the coordination-shell volume and its Wadell sphericity $\pi^{1/3}(6V)^{2/3}/A$, which equals one for a sphere and decreases for flattened or irregular shells.

Ligand electronegativity summaries. These features comprise the mean, standard deviation, minimum, maximum, sum, and count of the donor electronegativities on six scales (Pauling, Allen, Ghosh, Gunnarsson–Lundqvist, Miedema, and Robles–Bartolotti (Mentel, 2026)), together with the number of donors without an entry on each scale.

Spin state and oxidation state. The two regression tasks use three recorded site annotations: the spin quantum number, the spin per Fe atom, and the oxidation state.

Measured Mössbauer spectral parameters. The two classification tasks use the experimentally measured isomer shift and quadrupole splitting of the same Fe site.

Wavefunction-derived electronic summaries. The descriptors retained for Iron Spin State alone include the following pooled features: the molecular polarity index of the surface ESP (Lu & Chen, 2012b); the maximum, minimum, mean, standard deviation, and range of the ADCH charges (Lu & Chen, 2012a) of the coordinating atoms; dipole, quadrupole, and octupole moments of Fe alone, Fe with its coordinating atoms, the whole molecule, and the ligands alone; ESP statistics over the whole van der Waals surface and over the coordination-shell surface; the strongest ESP maxima and minima with their values and positions relative to Fe; and the lowest ESP-basin extrema with their values and distances from Fe.

D.4 Hand-crafted 3D descriptor baselines

We compare three representative hand-crafted 3D descriptors: SOAP, LMBTR, and GRIND.

Smooth Overlap of Atomic Positions (SOAP). SOAP represents local atomic environments through rotationally invariant power spectra of neighbor densities expanded in radial basis functions and spherical harmonics. We compute SOAP with DScribe (Himanen et al., 2020). On *QM9-sample*, the resulting atomic representations are averaged over the whole molecule, whereas on *MB* they are centered on the Fe coordination environment.

Local Many-Body Tensor Representation (LMBTR). LMBTR describes a local environment through distributions of inverse distances and angles around a center. We compute LMBTR with DScribe (Himanen et al., 2020), using the pairwise-distance term with exponential weighting. On *QM9-sample*, each atom serves as a center, whereas on *MB* the Fe site defines the center.

GRid-INdependent Descriptors (GRIND). GRIND summarizes molecular interaction fields through distance-dependent auto- and cross-correlograms of selected favorable grid nodes. We use an RDKit-based implementation (Landrum et al., 2026) with four probe fields: an electrostatic field from DFT-derived ADCH charges, two hydrogen-bond probes, and a steric probe. The fields are sampled on a molecule-centered grid, and the most favorable nodes from each field are retained.

For all three descriptors, PCA is fitted on the training split to retain 80% of the variance, and the fitted transformation is applied unchanged to the validation and test splits.

D.5 End-to-end DL baselines

All four DL baselines (SchNet, ViSNet, MolSpectra, and MABNet) use their official open-source implementations and respective training protocols, with task-specific training and evaluation performed on the same 1498-molecule *QM9-sample* split as the descriptor-based models. MolSpectra is pretrained on the pretraining dataset released by its authors before being fine-tuned on the *QM9-sample* training split.

E Implementation details

E.1 Source library instantiation

The source library $\mathcal{S}(\mathcal{M})$ contains $n_s = 13$ sources chosen to cover the chemical signals typically used to construct 3D descriptors. These include nuclear identities, electronegativities on six scales, and ADCH charges, each attached to atomic positions. The remaining sources comprise wavefunction-derived electrostatic-potential extrema on the molecular surface, electron-density extrema on the fragment Hirshfeld surface, and electrostatic-potential basin extrema in space. Each source is defined in Appendix H.1.

The source library $\mathcal{S}(\mathcal{M})$ is built from three types of input: molecular structures, element-wise lookup tables, and DFT wavefunctions. Atomic identities, represented by atomic numbers, are read directly from the molecular structures. The six electronegativity channels are populated using fixed element-wise lookup tables (Mentel, 2026). The atomic-dipole-corrected Hirshfeld (ADCH) charges (Lu & Chen, 2012a), whole-molecule surface ESP extrema, fragment-boundary electron-density extrema, and ESP-basin extrema are extracted from the wavefunctions with Multiwfn (Lu & Chen, 2012b; Lu, 2024). For *QM9-sample*, wavefunctions are computed on the standard QM9 geometries (Ramakrishnan et al., 2014) in Gaussian 09 at the B3LYP/6-31G(2df,p) level (Becke, 1993; Lee et al., 1988; Binkley et al., 1980; Frisch et al., 1984); for *MB*, geometries are optimized and wavefunctions are computed in Gaussian 16 at the PBE0-D3(BJ)/def2-SVPP level (Adamo & Barone, 1999; Weigend & Ahlrichs, 2005; Grimme et al., 2010; 2011).

Descriptors characterize the whole molecule for *QM9-sample* and the coordination environment of the Fe site for *MB*. To remove dependence on global orientation, all thirteen source clouds are expressed in a shared canonical frame intrinsic to each molecule. For *QM9-sample*, the origin is the geometric center of the heavy atoms, and the x axis points toward the heavy atom farthest from the origin. The y axis is the normalized projection, onto the plane perpendicular to x, of the vector from the origin to the heavy atom farthest from the x axis; $z = x \times y$. This construction defines a proper rotation and therefore preserves chirality. For molecules with fewer than three heavy atoms or with collinear heavy atoms, all atoms are used instead; if they remain collinear, an off-axis point from a non-atomic source is used to complete the frame. For *MB*, the canonical frame is aligned with the principal axes of the full molecule's nuclear-charge distribution, with axis directions fixed by the standard orientation returned by geometry optimization. Each Fe-site record is then centered on its target Fe atom.

E.2 Address instantiation

In *Pangaea*, each normalized address is $\theta = (u_b, z_1, \ldots, z_{14}) \in [0,1]^{15}$. The first coordinate u_b selects the branch, while the remaining coordinates control operator choices and their parameters within that branch. Each branch includes only its key design dimensions. The resulting per-branch coordinate counts are summarized in Table S6, and the following paragraphs detail how these co-

ordinates are allocated within each branch. Resolution, aggregation, and normalization settings are included among the parameters of their respective operators.

Table S6: Counts of address coordinates by branch, stage, and type. Slashes separate the maximum counts for different operator categories. Total counts the distinct coordinates allocated across all operators in a branch, excluding the branch selector u_b.

Branch	κ choice	κ params	τ choice	τ params	ψ choice	ψ params	Total
Point	1	1	1	7	1	3/2	14
Field	1	2	1	4/3	1	3/2/0	12
Expansion	1	0	1	6/5	1	3/2/1	12
Statistics	1	0	1	3/2	1	4/3/2	12

In the point branch, one coordinate selects κ, and one sets its support-realization parameter, since the point cloud itself is the carrier and its support can be relocated. τ uses one coordinate to select the operator and seven for its parameters. Of these, two set the resolution along the two matrix axes, and one controls the aggregation or normalization of point or event contributions. The roles of the remaining four depend on the selected operator. For projection, they control orientation and spatial extent. For local or manifold embedding, they control locality, weighting, and embedding scale or regularization. For event or interaction operators, they control geometric encoding, smoothing or ordering, and contribution weighting or mixing. ψ uses one coordinate to select the operator and three or two parameter coordinates. Two control spatial selection and weighting, or spectral and statistical transformations, depending on the operator. Operators with configurable pooling use a third parameter coordinate for the aggregation rule.

In the field branch, one coordinate selects κ. Two parameter coordinates set its field construction and spatial scale, since lifting a discrete source to a continuous field places a spatial profile at every source point. τ uses one coordinate to select the operator. All operators use two parameter coordinates for the resolution budgets. Map-valued operators use two further parameter coordinates for spatial extent and internal shape. Box-sampling operators use one further parameter coordinate for spatial extent, since the molecular bounding box fixes their shape. ψ uses one coordinate to select a compatible Compression operator, with three, two, or zero parameter coordinates. Matrix intermediates follow the point-branch allocation. For non-matrix intermediates, spectral and multiscale summaries use two coordinates for value or scale transformations and centering or mixing; pair histograms use two for distance range and the value exponent. Fixed summaries and vector finalization require no parameter coordinates. Non-matrix operators use fixed reduction rules.

In the expansion branch, one coordinate selects κ, with no additional Formation parameters. τ uses one coordinate to select the operator and six or five parameter coordinates. Three control the neighborhood cutoff, the treatment of channel weights, and an operator-specific shape or mixing setting. Two set the radial and degree resolutions of the feature tensor. Operators with adjustable radial scaling or bandwidth use one additional parameter coordinate. ψ uses one coordinate to select the operator. Compact summaries return 24 components and use three parameter coordinates. Two control value transformations, axis selection, or weighting and mixing, depending on the operator; the third selects the final normalization or value transformation. Full energy and Gram representations use two parameter coordinates for value exponents and centering or mixing. The raw-tensor option uses one parameter coordinate to mix the raw and normalized vectors.

In the statistics branch, one coordinate selects κ. It has no additional parameters, since events are enumerated on the source's own points, with no support to relocate and no kernel to choose. τ uses one coordinate to select the operator and three or two parameter coordinates. Two control the rule for deriving an event's weight from its participants' attributes (the weight mode) and the grouping of events into channels (the channel mode). The four operators that enumerate events within a neighborhood use a third for the cutoff radius. ψ uses one coordinate to select the operator. All operators use two parameter coordinates for channel pooling and vector normalization. Operators with adjustable smoothing or retention use two further coordinates for a width budget and the smoothing or retention setting. Operators controlled only by a bin count, output width, or moment order use one further coordinate. Fixed summary statistics require no additional parameters. These allocations give four, three, or two parameter coordinates in total.

The point branch requires the largest allocation of 14 coordinates. Reusing these positions across branches and adding the branch selector gives $d = 15$. Coordinates unused by the selected pipeline are inactive.

E.3 Closed-loop search protocol

IRA performs 2000 evaluations per *QM9-sample* task and 500 per *MB* task, beginning with 600 and 50 random warm-up trials, respectively, and proposes all remaining pipelines itself. Each run uses one of three random seeds (41–43), and the reported results are averaged over these seeds unless otherwise stated. After the search budget is exhausted, we fix the pipeline with the best search score. The learner is then refitted on the original training split using the baseline descriptors augmented with descriptors from the selected pipeline and evaluated on the held-out test set, as detailed in Appendix C. IRA is powered by `gpt-5.6-terra` with medium reasoning effort[1].

E.4 Black-box optimizers as search strategies

For the search strategy comparison, we select three representative black-box optimizers: TPE (Bergstra et al., 2011), CMA-ES (Hansen & Ostermeier, 2001), and PSO (Kennedy & Eberhart, 1995). All strategies are evaluated on *QM9-sample* with SVR as the downstream learner and the same budget of 2,000 evaluations per task.

Tree-structured Parzen Estimator (TPE). TPE proposes new addresses by fitting separate density estimators to better and worse previously evaluated addresses and favoring a large density ratio between them. We use two settings: vanilla TPE directly optimizes the full address vector, whereas TPE-struct respects the branch-conditional structure of the construction space, sampling only coordinates active in the selected branch and modeling operator choices and discrete operator parameters as categorical variables. In both settings, the first 600 proposals are random startup trials and the remaining 1400 are generated by the fitted TPE sampler.

Covariance Matrix Adaptation Evolution Strategy (CMA-ES). CMA-ES adapts a multivariate Gaussian search distribution from the outcomes of previously evaluated addresses. We use a population of 20 candidates, yielding 100 complete generations, an initial mean of $(0.5, \ldots, 0.5)$, and an initial step size of 0.35.

Particle Swarm Optimization (PSO). PSO updates each particle using its own best previously evaluated address and the best address found by the swarm. We use 40 particles and complete 2,000 evaluations through a sequential ask–tell loop, returning each evaluation result immediately rather than organizing evaluations into strictly synchronous generations.

E.5 Computational environment

Experiments were run on two Linux compute nodes: one with two Intel Xeon Gold 6326 CPUs and eight NVIDIA GeForce RTX 3090 GPUs, and the other with two Intel Xeon Gold 5318Y CPUs. Pipeline search and the ML learners ran on the CPUs; the DL baselines were trained on the RTX 3090 GPUs. Experiments used Python 3.10 with RDKit (2026.3.1), scikit-learn (1.7.2), XGBoost (3.1.2), PyTorch (2.10.0), Optuna (4.6.0), Nevergrad (1.0.12), and CMA (4.4.4).

F Additional results for the analysis of IRA

F.1 Search-strategy comparison

For the comparison of search strategies, we fix SVR as the downstream learner across the *QM9-sample* tasks and assign each strategy an evaluation budget of 2,000. Figure S1 shows the improvement over the *Topology & Global* baseline of the best-so-far search score as the evaluation budget increases. IRA's curves generally rise faster than those of the other search strategies, indicating more efficient use of the evaluation budget. The results are broadly consistent with those on the held-out test set in Figure 5, except for $\varepsilon_{\mathrm{HOMO}}$ and R^2. On these two tasks, IRA achieves larger

[1] `https://developers.openai.com/api/docs/models/gpt-5.6-terra`

gains in search score, but these gains do not translate into better held-out performance, which we hypothesize stems from weaker alignment between the search score and held-out performance.

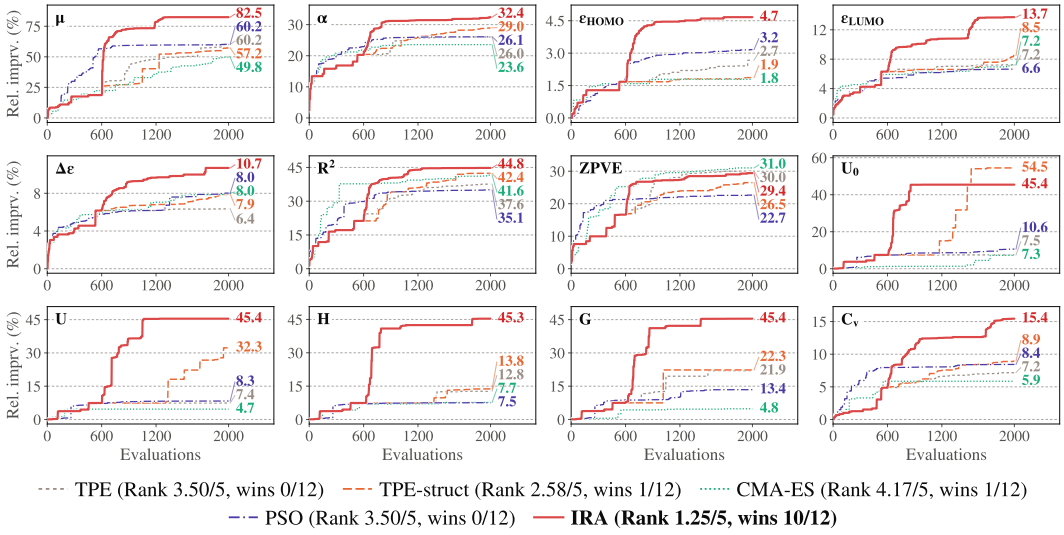


Figure S1: Search-score evolution of the incumbents in Figure 5. Each curve reports the incumbent's search-score improvement over the *Topology & Global* baseline.

F.2 Analysis of search guidance in IRA

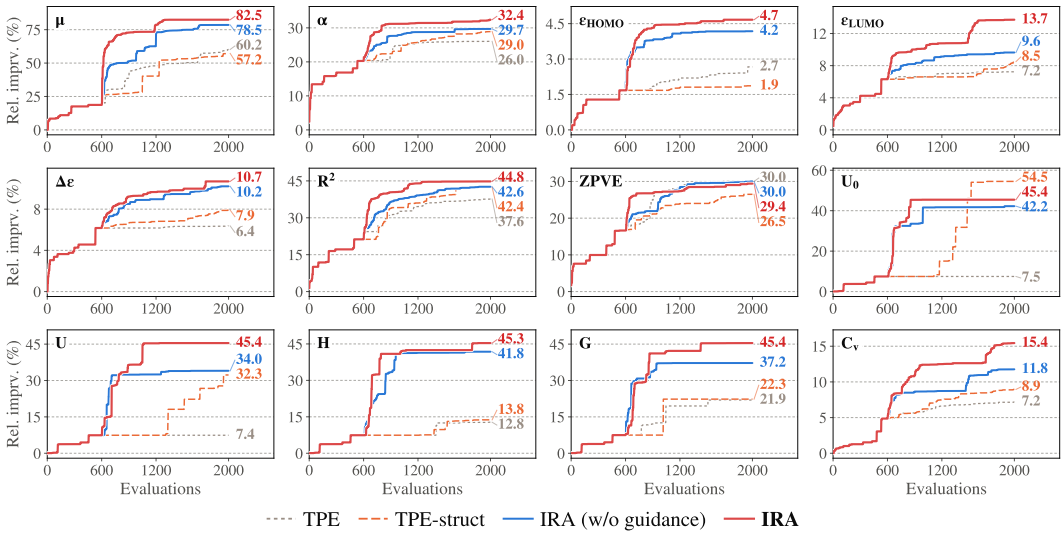


Figure S2: Per-task analysis of guidance from design-space semantics and chemical priors in IRA. Each curve reports the evolution of the incumbent's search-score improvement over the *Topology & Global* baseline.

For the search guidance analysis, we first construct IRA (w/o guidance) by anonymizing the information used to guide search. Specifically, all selectable components, including branches, operators, and other options, are replaced by indices. All continuous parameters are presented as unnamed values normalized to $[0, 1]$. Task information is removed entirely, except for the optimization direction. We compare IRA (w/o guidance) with TPE, TPE-struct and the IRA across the 12 *QM9-sample* tasks using SVR as the downstream learner, with all methods initialized from the same 600 random warm-up evaluations. Figure S2 reports the per-task results, showing that guidance improves IRA

on 11 of the 12 tasks and generally enables faster progress during search, demonstrating its benefit to both search effectiveness and efficiency. Even with all semantic and task-specific information anonymized, IRA (w/o guidance) still generally outperforms the black-box optimizers, suggesting that its underlying search strategy remains effective without explicit guidance.

F.3 Analysis of hierarchical revisions in IRA

We construct five variants to analyze the role of hierarchical revisions in IRA. For IRA (w/o hierarchy), we remove all control over the revision process and leave the choice of what to modify entirely to the LLM. For each of the other four variants, we retain only the remaining three revision levels and instruct the LLM to propose modifications within them. We evaluate these variants on six representative *QM9-sample* tasks using SVR as the downstream learner. For each task, we continue the evaluation from the seed on which the IRA performs best, with all settings initialized from the same 600 random warm-up evaluations. Figure S3 reports the results of 600 subsequent optimization steps for each method on each task. The results show that removing hierarchical revisions leads to worse search performance, with the w/o hierarchy variant finishing below IRA on five of the six tasks. Meanwhile, different tasks exhibit different dependencies on the revision levels: although removing a particular level can match IRA on some tasks, no incomplete hierarchy consistently preserves its performance across all six tasks, highlighting the complementary roles of the four revision levels.

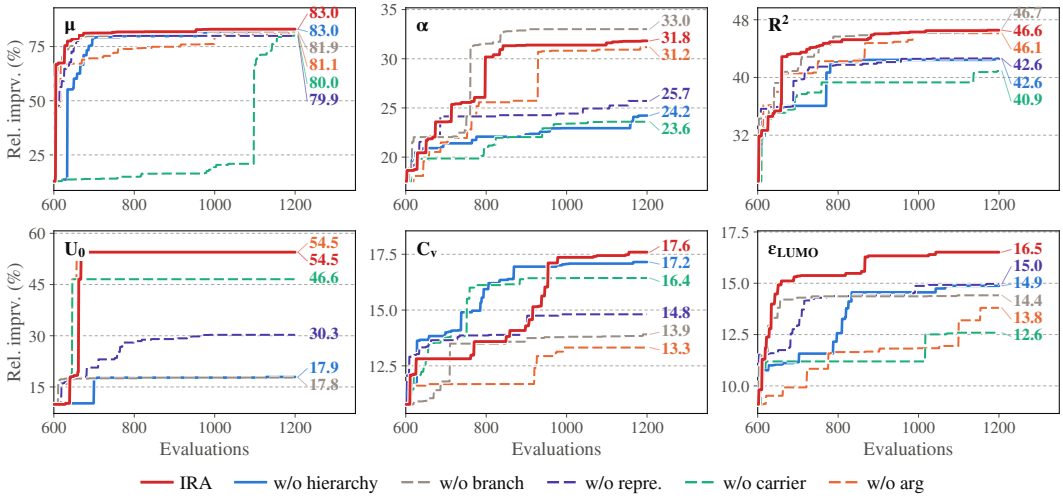


Figure S3: Per-task analysis of hierarchical revisions in IRA. Each curve reports the evolution of the incumbent's search-score improvement over the *Topology & Global* baseline.

G Pipelines designed by *Pangaea*

G.1 Designed pipelines for *QM9-sample*

Table S7 lists the final descriptor construction pipelines designed by *Pangaea* for the *QM9-sample* tasks. Each row corresponds to one learner–seed run and reports the selected operators κ, τ, and ψ, together with their effective parameters.

Table S7: Descriptor construction pipelines designed by *Pangaea* for *QM9-sample*.

Run	Branch	Formation (κ)	Extraction (τ)	Compression (ψ)
μ				
SVR 41	Field	Atomic-dipole-corrected Hirshfeld charge + Regularized Coulomb field `scale=0.13L`	Cylindrical integration `extent=2.48; inner=0.97 rho-max; grid=15x14`	Fourier moments `freq-cut=1.56; log-mix=0.75`
SVR 42	Field	Atomic-dipole-corrected Hirshfeld charge + Regularized Coulomb field `scale=0.09L`	Cylindrical integration `extent=2.42; inner=0.95 rho-max; grid=15x14`	Fourier moments `freq-cut=1.24; log-mix=0.55`

Table S7 (continued)

Run	Branch	Formation (κ)	Extraction (τ)	Compression (ψ)
SVR 43	Field	Atomic-dipole-corrected Hirshfeld charge + Regularized Coulomb field `scale=0.03L`	Cylindrical integration `extent=2.42; inner=0.95 rho-max; grid=4x11`	Fourier moments `freq-cut=3.19; log-mix=0.35`
Ridge 41	Field	Atomic-dipole-corrected Hirshfeld charge + Source-adaptive response `scale=0.21L`	Spherical integration `outer=2.42 rmax; inner=0.95 rmax; grid=15x15`	Fourier moments `freq-cut=2.21; log-mix=0.95`
Ridge 42	Field	Atomic-dipole-corrected Hirshfeld charge + Source-adaptive response `scale=0.18L`	Spherical integration `outer=2.42 rmax; inner=0.85 rmax; grid=15x15`	Fourier moments `freq-cut=3.35; log-mix=0.70`
Ridge 43	Field	Atomic-dipole-corrected Hirshfeld charge + Regularized Coulomb field `scale=0.01L`	Cylindrical integration `extent=2.42; inner=0.95 rho-max; grid=15x14`	Fourier moments `freq-cut=1.89; log-mix=0.85`
XGBoost 41	Field	Atomic-dipole-corrected Hirshfeld charge + Source-adaptive response `scale=0.18L`	Spherical integration `outer=2.42 rmax; inner=0.95 rmax; grid=6x13`	Fourier moments `freq-cut=3.19; log-mix=0.85`
XGBoost 42	Field	Atomic-dipole-corrected Hirshfeld charge + Regularized Coulomb field `scale=0.09L`	Cylindrical integration `extent=2.42; inner=0.95 rho-max; grid=12x14`	Fourier moments `freq-cut=2.86; log-mix=0.85`
XGBoost 43	Field	Atomic-dipole-corrected Hirshfeld charge + Regularized Coulomb field `scale=0.09L`	Cylindrical integration `extent=2.42; inner=0.95 rho-max; grid=15x12`	Fourier moments `freq-cut=3.19; log-mix=0.65`
α				
SVR 41	Point	Atomic number	Weighted interaction `power=4.85; sigma=0.10; three-body=0.45; triplet=0.75; grid=9x15; pool3D=min`	Fourier moments `freq-cut=1.40; log-mix=0.35`
SVR 42	Point	Atomic number	Weighted interaction `power=4.85; sigma=0.10; three-body=0.45; triplet=0.25; grid=12x15; pool3D=max`	Fourier moments `freq-cut=3.03; log-mix=0.85`
SVR 43	Point	Atomic number	Weighted interaction `power=2.75; sigma=0.05; three-body=0.95; triplet=0.05; grid=15x15; pool3D=sum`	Fourier moments `freq-cut=2.54; log-mix=0.05`
Ridge 41	Point	Atomic number	Weighted interaction `power=4.85; sigma=0.14; three-body=0.05; triplet=0.05; grid=9x15; pool3D=min`	Fourier moments `freq-cut=1.56; log-mix=0.05`
Ridge 42	Expansion	Gunnarsson–Lundqvist electronegativity	Shape and inertia `R=10; D=7; K=15; cutoff=0.58L if n>=15; transform=signed-log; rscale=1.99 rmax; moment-mix=0.45`	Axis moments `power=1.00; abs-mix=0.55; center-pool=mean; compression-norm=l2`
Ridge 43	Point	Atomic number	Weighted interaction `power=4.85; sigma=0.10; three-body=0.45; triplet=0.15; grid=13x15; pool3D=min`	Fourier moments `freq-cut=2.54; log-mix=0.05`
XGBoost 41	Point	Robles–Bartolotti electronegativity	Weighted interaction `power=6.25; sigma=0.07; three-body=0.65; triplet=0.65; grid=13x14; pool3D=median`	Multiscale summary `scale-exp=0.30; second-order=0.35`
XGBoost 42	Statistics	ESP-basin extrema	Triplet angle `cutoff=0.44L if n>=15; weight=product_scalar; channel=all`	Quantile–CDF `quantiles=14; pool=min/max; norm=l1`
XGBoost 43	Point	Ghosh electronegativity	Weighted interaction `power=4.85; sigma=0.14; three-body=0.25; triplet=0.95; grid=9x15; pool3D=median`	Column pooling `row-center=7.54; decay=2.55; pool2D=sum`
$\varepsilon_{\mathrm{HOMO}}$				
SVR 41	Point	Allen electronegativity	Weighted interaction `power=7.65; sigma=0.05; three-body=0.75; triplet=0.05; grid=13x12; pool3D=min`	Fourier moments `freq-cut=1.56; log-mix=0.75`
SVR 42	Field	Atomic-dipole-corrected Hirshfeld charge + Gaussian field density `scale=0.15L`	Voxel grid `margin=1.40; grid=12x12x10`	Multiscale summary `scale-exp=-0.70; second-order=0.78`
SVR 43	Field	Atomic-dipole-corrected Hirshfeld charge + Regularized Coulomb field `scale=0.06L`	Radial-shell map `outer=1.23 rmax; inner=0.05 rmax; grid=13x13`	PCA profile `angle=81.00 deg; window=0.73; pool2D=min`
Ridge 41	Point	Gunnarsson–Lundqvist electronegativity	Weighted interaction `power=4.85; sigma=0.16; three-body=0.35; triplet=0.05; grid=8x15; pool3D=min`	Fourier moments `freq-cut=2.86; log-mix=0.15`
Ridge 42	Point	Atomic-dipole-corrected Hirshfeld charge	Geometry-event histogram `inverse=0.35; sigma=0.08; lambda=1.60; triplet=0.55; grid=12x9; pool3D=median`	Column pooling `row-center=8.25; decay=1.65; pool2D=mean`
Ridge 43	Point	Pauling electronegativity	Weighted interaction `power=7.65; sigma=0.07; three-body=0.25; triplet=0.25; grid=9x6; pool3D=min`	Fourier moments `freq-cut=3.51; log-mix=0.95`
XGBoost 41	Field	Atomic-dipole-corrected Hirshfeld charge + Regularized Coulomb field `scale=0.12L`	Radial-shell map `outer=2.19 rmax; inner=0.09 rmax; grid=10x11`	Band selection `center=0.55 range; width=0.19; pool2D=mean`
XGBoost 42	Point	Atomic-dipole-corrected Hirshfeld charge `support=vdW surface`	Planar projection `az=234.00 deg; margin=1.48; elev=42.70 deg; roll=45.00 deg; grid=15x15; pool3D=sum`	Multiscale summary `scale-exp=-0.50; second-order=0.75`

Table S7 (continued)

Run	Branch	Formation (κ)	Extraction (τ)	Compression (ψ)
XGBoost 43	Point	Atomic-dipole-corrected Hirshfeld charge	Geometry-event histogram `inverse=0.55; sigma=0.07; lambda=4.60; triplet=0.45; grid=5x5; pool3D=mean`	Low-frequency compression `blur=2.44; blur-max=2.88; pool2D=sum`
$\varepsilon_{\mathrm{LUMO}}$				
SVR 41	Point	Gunnarsson–Lundqvist electronegativity	Weighted interaction `power=7.65; sigma=0.03; three-body=0.05; triplet=0.95; grid=15x15; pool3D=max`	Column pooling `row-center=10.50; decay=0.75; pool2D=sum`
SVR 42	Point	Ghosh electronegativity	Radial–angular density `rcut=1.06 rmax; sigma-r=0.12; sigma-a=0.11; mix=0.35; grid=14x10; pool3D=mean`	Polar compression `radial-angular=1.00; center-adapt=0.35; pool2D=min`
SVR 43	Point	Pauling electronegativity	Weighted interaction `power=7.65; sigma=0.04; three-body=0.95; triplet=0.95; grid=12x14; pool3D=min`	Column pooling `row-center=8.58; decay=1.35; pool2D=sum`
Ridge 41	Statistics	Atomic number	Triplet angle `cutoff=0.58L if n>=15; weight=unit; channel=element_pair`	Smoothed histogram `bins=7; softness=0.39; pool=channel_preserving; norm=maxabs`
Ridge 42	Statistics	Atomic number	Triplet angle `cutoff=0.58L if n>=15; weight=unit; channel=element_pair`	Smoothed histogram `bins=7; softness=0.39; pool=channel_preserving; norm=maxabs`
Ridge 43	Point	Miedema electronegativity	Geometry-event histogram `inverse=0.55; sigma=0.03; lambda=1.00; triplet=0.15; grid=11x12; pool3D=min`	Column pooling `row-center=4.39; decay=2.88; pool2D=sum`
XGBoost 41	Statistics	Pauling electronegativity	Pair distance `cutoff=0.53L if n>=15; weight=product_scalar; channel=all`	Quantile–CDF `quantiles=9; pool=median; norm=none`
XGBoost 42	Statistics	Gunnarsson–Lundqvist electronegativity	Radial density `cutoff=0.53L if n>=15; weight=abs_product_scalar; channel=all`	Quantile–CDF `quantiles=5; pool=mean+std; norm=none`
XGBoost 43	Statistics	Allen electronegativity	Pair distance `cutoff=0.58L if n>=15; weight=product_scalar; channel=all`	Quantile–CDF `quantiles=10; pool=median; norm=none`
$\Delta\varepsilon$				
SVR 41	Statistics	Atomic number	Triplet angle `cutoff=0.58L if n>=15; weight=unit; channel=neighbor_element_pair`	Smoothed histogram `bins=15; softness=0.39; pool=channel_preserving; norm=l2`
SVR 42	Point	Atomic number	Weighted interaction `power=2.75; sigma=0.03; three-body=0.65; triplet=0.55; grid=14x12; pool3D=median`	Fourier moments `freq-cut=2.21; log-mix=0.35`
SVR 43	Expansion	Molecular-surface ESP minima	Volumetric orthogonal `R=9; D=6; K=13; cutoff=0.13L if n>=15; transform=signed-log; rscale=1.54 rmax; damping=0.65`	Unfolding SVD `base-axis=3; power=0.88; center-pool=mean; compression-norm=l2`
Ridge 41	Statistics	Atomic number	Triplet angle `cutoff=0.33L if n>=15; weight=unit; channel=neighbor_element_pair`	Smoothed histogram `bins=15; softness=0.94; pool=channel_preserving; norm=none`
Ridge 42	Statistics	Pauling electronegativity	Triplet angle `cutoff=0.36L if n>=15; weight=mean_scalar; channel=all`	Quantile–CDF `quantiles=12; pool=channel_preserving; norm=l1`
Ridge 43	Statistics	Atomic number	Triplet angle `cutoff=0.43L if n>=15; weight=unit; channel=neighbor_element_pair`	Smoothed histogram `bins=15; softness=0.39; pool=channel_preserving; norm=l2`
XGBoost 41	Point	Allen electronegativity	Weighted interaction `power=7.65; sigma=0.05; three-body=0.95; triplet=0.95; grid=15x15; pool3D=min`	Fourier moments `freq-cut=3.84; log-mix=0.05`
XGBoost 42	Point	Pauling electronegativity	Weighted interaction `power=7.65; sigma=0.03; three-body=0.55; triplet=0.75; grid=14x13; pool3D=sum`	Column pooling `row-center=9.75; decay=2.85; pool2D=mean`
XGBoost 43	Point	Pauling electronegativity	Weighted interaction `power=5.90; sigma=0.03; three-body=0.55; triplet=0.85; grid=15x13; pool3D=max`	Column pooling `row-center=9.10; decay=1.95; pool2D=mean`
R^2				
SVR 41	Field	Pauling electronegativity + Regularized Coulomb field `scale=0.02L`	Spherical integration `outer=2.08 rmax; inner=0.35 rmax; grid=13x12`	Fourier moments `freq-cut=1.24; log-mix=0.05`
SVR 42	Field	Pauling electronegativity + Regularized Coulomb field `scale=0.02L`	Spherical integration `outer=2.08 rmax; inner=0.35 rmax; grid=13x12`	Fourier moments `freq-cut=0.91; log-mix=0.00`
SVR 43	Point	Pauling electronegativity	Weighted interaction `power=1.00; sigma=0.14; three-body=0.55; triplet=0.95; grid=14x7; pool3D=sum`	Fourier moments `freq-cut=3.19; log-mix=0.85`

Table S7 (continued)

Run	Branch	Formation (κ)	Extraction (τ)	Compression (ψ)
Ridge 41	Field	Pauling electronegativity + Regularized Coulomb field `scale=0.02L`	Spherical integration `outer=1.82 rmax; inner=0.40 rmax; grid=13x14`	Fourier moments `freq-cut=2.54; log-mix=0.05`
Ridge 42	Field	Pauling electronegativity + Regularized Coulomb field `scale=0.03L`	Spherical integration `outer=1.48 rmax; inner=0.75 rmax; grid=13x14`	Fourier moments `freq-cut=3.19; log-mix=0.05`
Ridge 43	Point	Pauling electronegativity	Weighted interaction `power=1.35; sigma=0.16; three-body=0.15; triplet=0.05; grid=6x7; pool3D=max`	Fourier moments `freq-cut=2.21; log-mix=0.95`
XGBoost 41	Field	Atomic number + Exponential field decay `scale=0.18L`	Spherical integration `outer=2.23 rmax; inner=0.25 rmax; grid=12x8`	Fourier moments `freq-cut=2.12; log-mix=0.75`
XGBoost 42	Point	Allen electronegativity	Radial–angular density `rcut=2.42 rmax; sigma-r=0.29; sigma-a=0.14; mix=0.95; grid=11x12; pool3D=max`	Band selection `center=0.00 range; width=0.20; pool2D=mean`
XGBoost 43	Field	Pauling electronegativity + Regularized Coulomb field `scale=0.02L`	Spherical integration `outer=2.25 rmax; inner=0.45 rmax; grid=7x14`	Multiscale summary `scale-exp=0.90; second-order=0.05`
ZPVE				
SVR 41	Point	Gunnarsson–Lundqvist electronegativity	Weighted interaction `power=6.25; sigma=0.12; three-body=0.15; triplet=0.38; grid=15x15; pool3D=min`	Fourier moments `freq-cut=0.91; log-mix=0.75`
SVR 42	Point	Atomic number `support=vdW surface`	Diffusion map `kappa=0.11L; tau=1.75; clip=[0.06,0.94]; attr=0.10; grid=15x10; pool3D=sum`	PCA profile `angle=72.00 deg; window=1.93; pool2D=sum`
SVR 43	Point	Atomic number	Weighted interaction `power=6.95; sigma=0.14; three-body=0.05; triplet=0.35; grid=15x15; pool3D=max`	Multiscale summary `scale-exp=0.10; second-order=0.75`
Ridge 41	Point	Gunnarsson–Lundqvist electronegativity	Weighted interaction `power=6.25; sigma=0.12; three-body=0.05; triplet=0.95; grid=7x15; pool3D=min`	Fourier moments `freq-cut=0.91; log-mix=0.95`
Ridge 42	Point	Atomic number `support=vdW surface`	Planar projection `az=234.00 deg; margin=1.40; elev=6.80 deg; roll=45.00 deg; grid=12x14; pool3D=sum`	Multiscale summary `scale-exp=-0.30; second-order=0.02`
Ridge 43	Point	Atomic number	Weighted interaction `power=7.65; sigma=0.05; three-body=0.45; triplet=0.95; grid=13x11; pool3D=min`	Fourier moments `freq-cut=3.19; log-mix=0.65`
XGBoost 41	Point	Gunnarsson–Lundqvist electronegativity	Weighted interaction `power=4.15; sigma=0.14; three-body=0.65; triplet=0.55; grid=7x5; pool3D=min`	Fourier moments `freq-cut=1.40; log-mix=0.25`
XGBoost 42	Point	Robles–Bartolotti electronegativity `support=vdW surface`	Smoothed local PCA `sigma=0.05L; window=0.71 rcut; rcut=3.43 sigma; exponent=0.75; grid=4x5; pool3D=mean`	Column pooling `row-center=1.26; decay=1.95; pool2D=sum`
XGBoost 43	Point	Robles–Bartolotti electronegativity `support=vdW surface`	Smoothed local PCA `sigma=0.13L; window=2.60 rcut; rcut=2.03 sigma; exponent=0.40; grid=4x11; pool3D=mean`	Low-frequency compression `blur=1.59; blur-max=1.88; pool2D=sum`
U_0				
SVR 41	Point	Atomic number	Local PCA `r=0.09L; window=1.65r; sigma=0.60r; decay=0.90; grid=15x9; pool3D=sum`	Multiscale summary `scale-exp=0.90; second-order=0.85`
SVR 42	Point	Atomic number	Local PCA `r=0.11L; window=1.92r; sigma=0.60r; decay=0.90; grid=6x12; pool3D=sum`	Multiscale summary `scale-exp=-0.20; second-order=0.40`
SVR 43	Statistics	Atomic number	Local shape `cutoff=0.43L if n>=15; weight=unit; channel=center_element`	Summary statistics `8 stats/channel; pool=channel_preserving; norm=none`
Ridge 41	Point	Atomic number	Local PCA `r=0.05L; window=1.79r; sigma=0.64r; decay=1.35; grid=13x9; pool3D=sum`	Multiscale summary `scale-exp=-0.30; second-order=0.75`
Ridge 42	Point	Atomic number	Local PCA `r=0.03L; window=1.79r; sigma=0.56r; decay=1.05; grid=8x11; pool3D=sum`	Multiscale summary `scale-exp=-0.50; second-order=0.45`
Ridge 43	Point	Fragment-surface electron-density maxima	Spherical projection `az=54.00 deg; shell=1.12 rmax; r0=q46; tilt=6.00 deg; grid=15x14; pool3D=min`	Directional profile `angle=187.20 deg; window=1.63; pool2D=sum`
XGBoost 41	Statistics	Ghosh electronegativity	Local shape `cutoff=0.43L if n>=15; weight=mean_scalar; channel=all`	Quantile–CDF `quantiles=14; pool=min/max; norm=maxabs`
XGBoost 42	Statistics	Atomic number	Triplet angle `cutoff=0.13L if n>=15; weight=unit; channel=element_pair`	Smoothed histogram `bins=12; softness=1.76; pool=channel_preserving; norm=maxabs`
XGBoost 43	Statistics	Atomic number	Local shape `cutoff=0.18L if n>=15; weight=unit; channel=neighbor_element_pair`	Channel-preserving bag `bag=12; pool=channel_preserving; norm=maxabs`

Table S7 (continued)

Run	Branch	Formation (κ)	Extraction (τ)	Compression (ψ)
U				
SVR 41	Point	Atomic number	Local PCA `r=0.05L; window=2.06r; sigma=0.64r; decay=0.75; grid=14x11; pool3D=sum`	Multiscale summary `scale-exp=-0.10; second-order=0.35`
SVR 42	Point	Atomic number	Local PCA `r=0.05L; window=1.92r; sigma=0.68r; decay=1.50; grid=8x11; pool3D=sum`	Multiscale summary `scale-exp=-0.50; second-order=0.05`
SVR 43	Statistics	Atomic number	Local shape `cutoff=0.42L if n>=15; weight=unit; channel=neighbor_element_pair`	Summary statistics `8 stats/channel; pool=channel_preserving; norm=none`
Ridge 41	Point	Atomic number	Local PCA `r=0.09L; window=1.52r; sigma=0.64r; decay=0.75; grid=5x13; pool3D=sum`	Multiscale summary `scale-exp=0.50; second-order=0.35`
Ridge 42	Statistics	Atomic number	Local shape `cutoff=0.33L if n>=15; weight=unit; channel=center_element`	Summary statistics `8 stats/channel; pool=channel_preserving; norm=none`
Ridge 43	Point	Fragment-surface electron-density maxima	Spherical projection `az=72.00 deg; shell=1.71 rmax; r0=q38; tilt=27.00 deg; grid=15x14; pool3D=max`	Directional profile `angle=126.00 deg; window=1.48; pool2D=sum`
XGBoost 41	Statistics	Allen electronegativity	Local shape `cutoff=0.43L if n>=15; weight=product_scalar; channel=all`	Quantile–CDF `quantiles=14; pool=mean+std; norm=maxabs`
XGBoost 42	Statistics	Atomic number	Local shape `cutoff=0.48L if n>=15; weight=unit; channel=element_pair`	Smoothed histogram `bins=12; softness=1.79; pool=channel_preserving; norm=maxabs`
XGBoost 43	Statistics	Atomic number	Pair distance `cutoff=0.21L if n>=15; weight=unit; channel=element_pair`	Channel-preserving bag `bag=12; pool=channel_preserving; norm=maxabs`
H				
SVR 41	Point	Atomic number	Local PCA `r=0.05L; window=1.38r; sigma=0.60r; decay=0.90; grid=15x13; pool3D=sum`	Multiscale summary `scale-exp=-0.90; second-order=0.85`
SVR 42	Point	Atomic number	Local PCA `r=0.06L; window=1.79r; sigma=0.72r; decay=1.65; grid=12x14; pool3D=sum`	Multiscale summary `scale-exp=-0.10; second-order=0.70`
SVR 43	Statistics	Atomic number	Local shape `cutoff=0.38L if n>=15; weight=unit; channel=element_pair`	Summary statistics `8 stats/channel; pool=channel_preserving; norm=none`
Ridge 41	Point	Atomic number	Smoothed local PCA `sigma=0.02L; window=1.79 rcut; rcut=3.08 sigma; exponent=0.27; grid=11x9; pool3D=sum`	Multiscale summary `scale-exp=-0.30; second-order=0.45`
Ridge 42	Point	Atomic number	Local PCA `r=0.01L; window=2.33r; sigma=0.60r; decay=1.95; grid=10x10; pool3D=sum`	Multiscale summary `scale-exp=-0.40; second-order=0.90`
Ridge 43	Point	Fragment-surface electron-density maxima	Spherical projection `az=126.00 deg; shell=1.32 rmax; r0=q38; tilt=15.00 deg; grid=15x14; pool3D=min`	Directional profile `angle=270.00 deg; window=1.93; pool2D=sum`
XGBoost 41	Statistics	Atomic number	Radial density `cutoff=0.28L if n>=15; weight=unit; channel=center_element`	Quantile–CDF `quantiles=14; pool=min/max; norm=maxabs`
XGBoost 42	Statistics	Atomic number	Local shape `cutoff=0.46L if n>=15; weight=unit; channel=element_pair`	Smoothed histogram `bins=12; softness=2.04; pool=channel_preserving; norm=maxabs`
XGBoost 43	Statistics	Atomic number	Local shape `cutoff=0.18L if n>=15; weight=unit; channel=neighbor_element_pair`	Channel-preserving bag `bag=12; pool=channel_preserving; norm=maxabs`
G				
SVR 41	Point	Atomic number	Smoothed local PCA `sigma=0.02L; window=2.06 rcut; rcut=3.43 sigma; exponent=0.79; grid=9x14; pool3D=sum`	Multiscale summary `scale-exp=0.50; second-order=0.70`
SVR 42	Point	Atomic number	Local PCA `r=0.06L; window=1.52r; sigma=0.80r; decay=2.25; grid=6x14; pool3D=sum`	Multiscale summary `scale-exp=0.50; second-order=0.25`
SVR 43	Statistics	Atomic number	Local shape `cutoff=0.42L if n>=15; weight=unit; channel=neighbor_element_pair`	Summary statistics `8 stats/channel; pool=channel_preserving; norm=none`
Ridge 41	Point	Atomic number	Local PCA `r=0.04L; window=1.65r; sigma=0.60r; decay=2.40; grid=6x12; pool3D=sum`	Multiscale summary `scale-exp=-0.90; second-order=0.20`
Ridge 42	Point	Atomic number	Local PCA `r=0.03L; window=1.65r; sigma=0.60r; decay=2.10; grid=13x15; pool3D=sum`	Multiscale summary `scale-exp=-0.56; second-order=0.22`
Ridge 43	Point	Atomic number	Smoothed local PCA `sigma=0.02L; window=2.06 rcut; rcut=3.43 sigma; exponent=0.66; grid=14x14; pool3D=sum`	Multiscale summary `scale-exp=0.10; second-order=0.45`

Table S7 (continued)

Run	Branch	Formation (κ)	Extraction (τ)	Compression (ψ)
XGBoost 41	Expansion	Atomic number	Pair-distance spectral R=9; D=7; K=15; cutoff=0.58L if n>=15; transform=signed-root; phase-mix=0.55	Channel-pair power power=2.50; centering=0.25; center-pool=mean; split-PCA=80.00pct variance
XGBoost 42	Statistics	Atomic number	Radial density cutoff=0.18L if n>=15; weight=unit; channel=neighbor_element_pair	Hard histogram bins=12; pool=channel_preserving; norm=none
XGBoost 43	Statistics	Atomic number	Triplet angle cutoff=0.28L if n>=15; weight=unit; channel=neighbor_element_pair	Channel-preserving bag bag=12; pool=channel_preserving; norm=maxabs
c_v				
SVR 41	Point	Pauling electronegativity	Weighted interaction power=7.65; sigma=0.19; three-body=0.30; triplet=0.05; grid=6x7; pool3D=min	Fourier moments freq-cut=1.89; log-mix=0.05
SVR 42	Point	Pauling electronegativity	Weighted interaction power=6.25; sigma=0.14; three-body=0.55; triplet=0.65; grid=12x13; pool3D=sum	Fourier moments freq-cut=2.86; log-mix=0.05
SVR 43	Point	Pauling electronegativity	Weighted interaction power=7.65; sigma=0.05; three-body=0.35; triplet=0.55; grid=15x15; pool3D=max	PCA profile angle=90.00 deg; window=0.73; pool2D=mean
Ridge 41	Point	Pauling electronegativity	Weighted interaction power=7.65; sigma=0.17; three-body=0.25; triplet=0.01; grid=7x9; pool3D=min	Fourier moments freq-cut=3.19; log-mix=0.02
Ridge 42	Field	Pauling electronegativity + Exponential field decay scale=0.06L	Spherical integration outer=1.40 rmax; inner=0.35 rmax; grid=12x10	Fourier moments freq-cut=2.86; log-mix=0.35
Ridge 43	Point	Allen electronegativity	Weighted interaction power=4.15; sigma=0.14; three-body=0.85; triplet=0.35; grid=11x6; pool3D=max	Fourier moments freq-cut=2.54; log-mix=0.95
XGBoost 41	Expansion	Allen electronegativity	Shape and inertia R=10; D=8; K=17; cutoff=0.58L if n>=15; transform=signed-log; rscale=0.86 rmax; moment-mix=0.55	Raw tensor L2-mix=0.85; center-pool=mean; split-PCA=80.00pct variance
XGBoost 42	Statistics	Pauling electronegativity	Pair distance cutoff=0.53L if n>=15; weight=mean_scalar; channel=all	Quantile–CDF quantiles=14; pool=mean+std; norm=none
XGBoost 43	Point	Allen electronegativity	Geometry-event histogram inverse=0.85; sigma=0.05; lambda=2.20; triplet=0.85; grid=15x12; pool3D=max	Band selection center=0.95 range; width=0.09; pool2D=mean

Continuous parameters are rounded to two decimal places; L denotes the bounding-box diagonal of the point cloud (L_X in Appendix H.3).

G.2 Visual analysis of the pipelines designed by *Pangaea* for *MB*

Figure S4 visualizes the pipelines designed by *Pangaea* for the four *MB* tasks. Across the four *MB* tasks, the designed pipelines produce descriptors that recover spatial distributions around Fe that are closely related to the chemical basis of the targets. For δ, which reflects the s-electron density at the Fe nucleus and is governed by electronic interactions between the ligands and the Fe center, the designed pipeline uses electron-density extrema around Fe and their pairwise distances to characterize the morphology of the ligand electron density. This representation captures whether the density is concentrated in a single cluster or distributed more evenly, as well as how tightly it accumulates around Fe, providing information related to the s-electron density at the nucleus. For ΔE_Q, which reflects the electric field gradient at the nucleus and depends on how the coordination shell deviates from cubic symmetry, the designed pipeline uses electronegativity-weighted pair distances to characterize the local coordination geometry. For the iron oxidation state, which is governed by the bond valence sum and depends jointly on bond lengths and the identities of the coordinating atoms, the designed pipeline yields an ordered profile of atom types and their distances from Fe. For the iron spin state, although the designed pipeline omits electronegativity weighting, its descriptor captures the size, asphericity, and anisotropy of the coordination shell as viewed from each coordinating atom. This coordination-shell shape information distinguishes spin states because the shell expands and becomes more irregular as the antibonding e_g/σ^* orbitals are populated, so its size and regularity track whether the Fe center is in a low- or high-spin state.

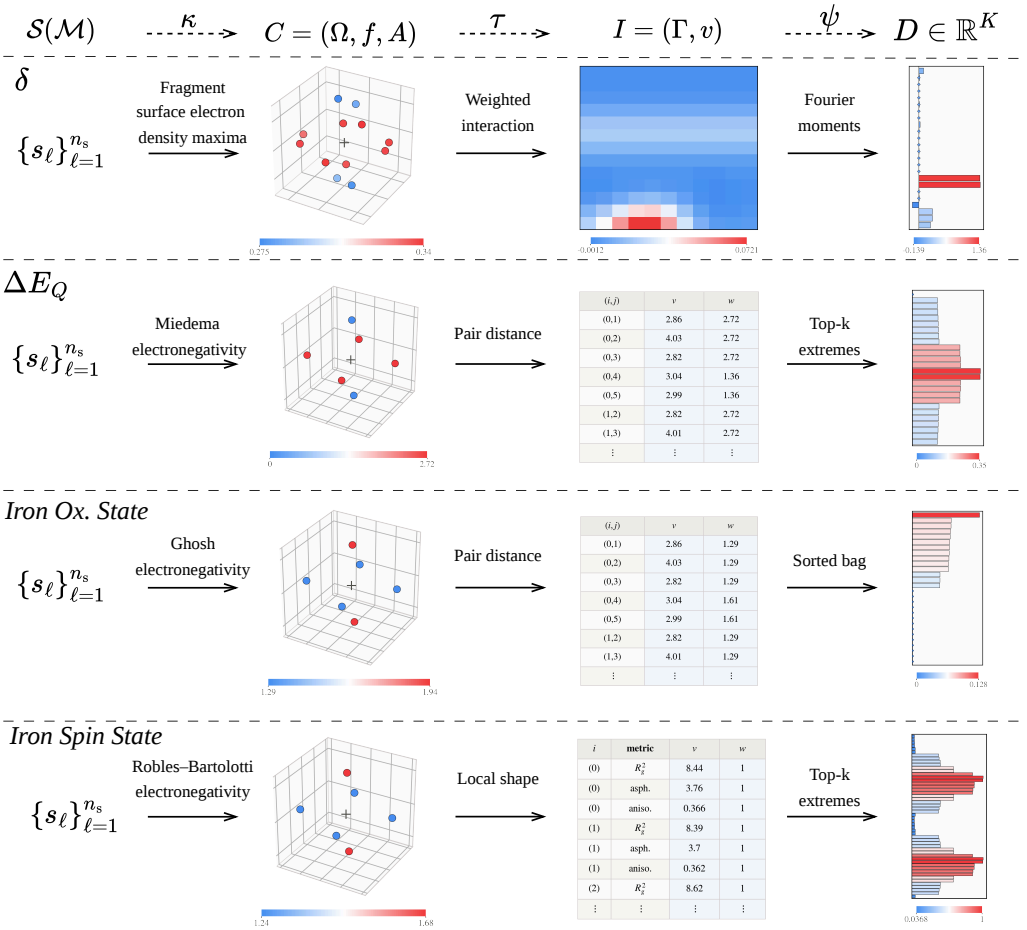


Figure S4: Pipelines designed by *Pangaea* for the four *MB* tasks, illustrated for a single Fe site.

H Details of the Unified Construction Space

This appendix gives a complete definition of the Unified Construction Space. We first specify the shared source library (Appendix H.1), together with the branch selector and the canonical address layout within each branch (Appendix H.2). Before detailing the branches, we introduce the common pipeline notation and parameter conventions (Appendix H.3). We then define each branch in turn (Appendices H.4–H.7), giving its concrete implementation and address decoding.

H.1 Shared source library

All four branches draw from the shared source library $\mathcal{S}(\mathcal{M}) = \{s_\ell = (X_\ell, a_\ell)\}_{\ell=1}^{n_s}$, instantiated with $n_s = 13$ attributed point clouds (Table S8). For each source s_ℓ, $X_\ell = (\mathbf{x}_{\ell,i})_{i=1}^{N_\ell} \in \mathbb{R}^{N_\ell \times 3}$ collects the coordinates of its native spatial support, and $a_\ell = (a_{\ell,i})_{i=1}^{N_\ell} \in \mathbb{R}^{N_\ell}$ records the chemical signal attached to each support point as a scalar attribute.

Table S8: Details of the shared source library.

ℓ	Source signal	Point locations X_ℓ	Stored scalar $a_{\ell,i}$
1	Atomic number	X_{atom}	$\log(1 + Z_i)$
2	Pauling electronegativity	X_{atom}	$\widehat{\chi}_i^{(\text{Pauling})}$
3	Allen electronegativity	X_{atom}	$\widehat{\chi}_i^{(\text{Allen})}$
4	Ghosh electronegativity	X_{atom}	$\widehat{\chi}_i^{(\text{Ghosh})}$

Table S8 (continued)

ℓ	Source signal	Point locations X_ℓ	Stored scalar $a_{\ell,i}$
5	Gunnarsson–Lundqvist electronegativity	X_{atom}	$\widehat{\chi}_i^{(\mathrm{GL})}$
6	Miedema electronegativity	X_{atom}	$\widehat{\chi}_i^{(\mathrm{Miedema})}$
7	Robles–Bartolotti electronegativity	X_{atom}	$\widehat{\chi}_i^{(\mathrm{RB})}$
8	Atomic-dipole-corrected Hirshfeld charge	X_{atom}	q_i^{ADCH}
9	Molecular-surface ESP maxima	Local maxima of V_{ESP} on S_ρ	$V_{\mathrm{ESP}}(\mathbf{x}_{9,i})$
10	Molecular-surface ESP minima	Local minima of V_{ESP} on S_ρ	$V_{\mathrm{ESP}}(\mathbf{x}_{10,i})$
11	Fragment-surface electron-density maxima	Local maxima of ρ on S_{H}	$\rho(\mathbf{x}_{11,i})$
12	Fragment-surface electron-density minima	Local minima of ρ on S_{H}	$\rho(\mathbf{x}_{12,i})$
13	ESP-basin extrema	Basin extrema of V_{ESP}	Signed-log $V_{\mathrm{ESP}}(\mathbf{x}_{13,i})$

Sources 1–8 share the atomic-coordinate support X_{atom}. For *QM9-sample*, X_{atom} comprises all atoms in the molecule; for *MB*, it comprises the atoms in the Fe-centered coordination environment. For atom i, let Z_i denote its atomic number; source 1 stores $a_{1,i} = \log(1 + Z_i)$. Sources 2–7 use six tabulated electronegativity conventions (Mentel, 2026), with $\chi_i^{(\ell)}$ denoting the electronegativity of element Z_i on the scale of source ℓ. To place these conventions on comparable numerical scales while keeping preprocessing independent of the molecular datasets, each is standardized using the finite entries of its elemental reference table: $\widehat{\chi}_i^{(\ell)} = (\chi_i^{(\ell)} - \overline{\chi}_{\mathrm{ref}}^{(\ell)})/\,\mathrm{sd}_{\mathrm{ref}}^{(\ell)}$, where $\overline{\chi}_{\mathrm{ref}}^{(\ell)}$ and $\mathrm{sd}_{\mathrm{ref}}^{(\ell)}$ are the population mean and standard deviation (SD) of those entries. Finally, let q_i^{ADCH} denote the atomic-dipole-corrected Hirshfeld charge of atom i in units of the elementary charge. Source 8 stores $a_{8,i} = q_i^{\mathrm{ADCH}}$ directly, retaining the physical zero and sign of the charge.[2] Missing electronegativity entries and unavailable ADCH values are represented by zero, corresponding respectively to the reference-table mean and neutral charge.

Sources 9 and 10 encode the local maxima and minima, respectively, of the total electrostatic potential on a full-molecule electron-density isosurface. Let ρ denote the electron density and define this surface as $S_\rho = \{\mathbf{x} \in \mathbb{R}^3 : \rho(\mathbf{x}) = 10^{-3}\ \text{a.u.}\}$. Restricting V_{ESP} to S_ρ, we take its maximum locations $\mathbf{x}_{9,i}$ and minimum locations $\mathbf{x}_{10,i}$ as the rows of X_9 and X_{10}, respectively. The corresponding scalar attributes are $a_{9,i} = V_{\mathrm{ESP}}(\mathbf{x}_{9,i})$ and $a_{10,i} = V_{\mathrm{ESP}}(\mathbf{x}_{10,i})$.

Sources 11 and 12 encode the local maxima and minima, respectively, of the electron density on a fragment Hirshfeld surface S_{H} (Spackman & Byrom, 1997). For *QM9-sample*, the fragment comprises all non-hydrogen atoms and its environment comprises the hydrogen atoms. For *MB*, the fragment comprises the target Fe atom and its coordinating atoms, while its environment comprises the remaining ligand atoms. Restricting ρ to the resulting S_{H}, we take its maximum locations $\mathbf{x}_{11,i}$ and minimum locations $\mathbf{x}_{12,i}$ as the rows of X_{11} and X_{12}, respectively. The corresponding scalar attributes are $a_{11,i} = \rho(\mathbf{x}_{11,i})$ and $a_{12,i} = \rho(\mathbf{x}_{12,i})$.

Source 13 encodes the positive-region attractors and negative-region repulsors of the total electrostatic potential over three-dimensional space. We obtain their locations $\mathbf{x}_{13,i}$ from an ESP basin analysis (Lu & Chen, 2012b) and use them as the rows of X_{13}. To accommodate the broad dynamic range of the corresponding ESP values, the source stores the signed-log-transformed scalar $a_{13,i} = \mathrm{sgn}(V_{\mathrm{ESP}}(\mathbf{x}_{13,i}))\log(1 + |V_{\mathrm{ESP}}(\mathbf{x}_{13,i})|)$. This transformation limits the influence of large-magnitude extrema while preserving the ESP sign.

H.2 Branch selection and address layout

Every pipeline in the library is controlled through the normalized address $\theta = (u_b, z_1, \ldots, z_{14})$. The first coordinate selects the branch, $b(u_b) = (\mathrm{pt}, \mathrm{fld}, \mathrm{exp}, \mathrm{stat})_{1+\min\{\lfloor 4u_b \rfloor, 3\}}$; the remaining fourteen are decoded within the selected branch, and Table S9 records what each of them controls.[3]

H.3 Common pipeline notation

In the pipeline of Section 4, the carrier is $C = (\Omega, \mathbf{f}; \mathcal{A})$, where $\Omega \subseteq \mathbb{R}^3$ is either a finite spatial support or a continuous spatial domain, and $\mathbf{f} : \Omega \to \mathbb{R}^{c_C}$ is a scalar or finite-channel signal on that

[2] The computational procedures for Sources 8–13 are detailed in Appendix E.1.

[3] All selection menus use one-based indices in the order listed.

Table S9: Branch-specific operator choices and parameters encoded by the address coordinates.

Address coordinate	Point	Field	Expansion	Statistics
z_1	κ choice (source)	κ choice (source)	κ choice (source)	κ choice (source)
z_2	support realization	field construction	–	–
z_3	τ choice	τ choice	τ choice	τ choice
z_4	ϑ_m	kernel scale	ϑ_m	ϑ_m
z_5	β_m	ϑ_m	β_m	β_m
z_6	γ_m	β_m	γ_m	γ_m
z_7	δ_m	–	δ_m	–
z_8, z_9	H, W	H_0, W_0	N_r, N_d	$N_{\mathrm{hist}}, N_{\mathrm{bag}}$
z_{10}	τ aggregation	–	–	channel pooling
z_{11}	ψ choice	ψ choice	ψ choice	ψ choice
z_{12}, z_{13}	λ_m, ς_m	λ_m, ς_m	λ_m, ς_m	operator parameters
z_{14}	ψ aggregation	matrix ψ aggregation	Compression finalizer	vector normalization

support. The optional finite auxiliary structure $\mathcal{A}$ is retained only when required by a typed downstream operator. The Extraction operator τ produces the intermediate representation $I = (\Gamma, \mathbf{v})$, where Γ is a finite index structure comprising indices and any ordering, adjacency, coordinate, or channel records required by the Compression operator, and $\mathbf{v} : \Gamma \to \mathbb{R}^{c_I}$ assigns c_I values to each index. The Compression operator ψ encodes this representation into $D \in \mathbb{R}^{K_\varphi}$, with pooling and normalization where needed.

Throughout these branch definitions, the τ-stage parameters of an operator labeled by m are denoted by $\vartheta_m, \beta_m, \gamma_m$, and δ_m, and its ψ-stage parameters by λ_m and ς_m, with unused components omitted. When present, selectable aggregation rules are denoted by ϱ_m at the τ stage, while ζ_m denotes a selectable aggregation or finalization rule at the ψ stage. ν_m denotes event-count normalization for soft-histogram τ operators, and υ_m denotes final-vector normalization when present.[4] Throughout this section, $\epsilon = 10^{-12}$ is the fixed numerical stabilizer, $\mathbf{1}$ an all-ones object of context-determined size, $\mathbb{I}[\cdot]$ an indicator, and L_X the bounding-box diagonal of a point set X.[5]

H.4 Point branch

H.4.1 Branch overview

A point-branch pipeline takes the following form:

$$\mathcal{S}(\mathcal{M}) \xrightarrow{\kappa_{\mathrm{pt}}} C_{\mathrm{pt}} = (X, a; \varnothing) \xrightarrow{\tau_{\mathrm{pt}}} I_{\mathrm{pt}} \in \mathbb{R}^{H \times W} \xrightarrow{\psi_{\mathrm{pt}}} D \in \mathbb{R}^K,$$

where $X \in \mathbb{R}^{N \times 3}$ contains the 3D coordinates and $a \in \mathbb{R}^N$ contains one scalar attribute per point.

H.4.2 Point-branch Formation (κ_{pt})

For each source $s_\ell = (X_\ell, a_\ell)$ in the shared library, we define a corresponding Formation operator κ_{pt} that selects this source and constructs an attributed point cloud carrying its physical signal. Its support realization u_κ is selected from the source's own points (dir), the van der Waals surface (surf), and the van der Waals volume (vol). Direct realization leaves the source unchanged, $X = X_\ell$, and is available to all thirteen sources. Surface and volume realizations resample it and are restricted to the first eight atom-supported sources in Table S8.

For these sources, let $\mathbf{x}_i$ and r_i^{vdW} denote the position and van der Waals radius of atom i. Surface support is generated by placing quasi-uniform Fibonacci points on each atomic sphere and retaining those outside all other spheres. Volume support consists of points from a Halton sequence that lie

[4] These symbols, together with any frame vectors or intermediate quantities introduced for a concrete operator, are local to operator m and imply neither shared values nor shared meanings across operators. When the operator is fixed, as within a registry or decoding table and the text around it, we omit the label m for clarity.

[5] Once the parameters, resolution variables, and aggregation rules of an operator have been stated, we treat them as part of the operator and drop them from its argument list.

inside the union of the van der Waals balls. Each point p is assigned an owner $o(p)$: its generating atom on the surface, or the containing atom that minimizes $\|p - \mathbf{x}_i\|_2^2/(r_i^{\mathrm{vdW}})^2$ in the volume.

The candidate points are stratified by owner and deterministically reduced to a geometry-dependent target budget clipped between 128 and 256 points; the retained coordinates form X. Let n_i be the number of retained points owned by atom i. A retained point p with $o(p) = i$ receives quadrature weight $w_p^{\mathrm{quad}} = A_i^{\mathrm{exp}}/n_i$ on the surface or $w_p^{\mathrm{quad}} = V_i^{\mathrm{own}}/n_i$ in the volume, where A_i^{exp} is the exposed area of atom i and V_i^{own} is the volume assigned to it by the ownership rule. The carrier attributes are unified as

$$a_p^{(\ell,u_\kappa)} = \begin{cases} a_{\ell,p}, & u_\kappa = \mathrm{dir}, \\ a_{\ell,o(p)} w_p^{\mathrm{quad}}, & u_\kappa \in \{\mathrm{surf}, \mathrm{vol}\}, \quad 1 \le \ell \le 7, \\ q_{o(p)}^{\mathrm{ADCH}} w_p^{\mathrm{quad}} \big(4\pi (r_{o(p)}^{\mathrm{vdW}})^2\big)^{-1}, & u_\kappa = \mathrm{surf}, \quad \ell = 8, \\ q_{o(p)}^{\mathrm{ADCH}} w_p^{\mathrm{quad}} \big(4\pi (r_{o(p)}^{\mathrm{vdW}})^3/3\big)^{-1}, & u_\kappa = \mathrm{vol}, \quad \ell = 8. \end{cases}$$

H.4.3 Point-branch Extraction (τ_{pt})

The Extraction operator τ_{pt} maps the attributed point cloud $C_{\mathrm{pt}} = (X, a; \varnothing)$ to $I_{\mathrm{pt}} \in \mathbb{R}^{H \times W}$ by first constructing coordinates and associated values or a pairwise matrix from the cloud and then rasterizing the result. Table S10 specifies the twelve operator-specific constructions, and the rules shared across operators are defined after the table.

In the table, $r_i = \|\mathbf{x}_i\|_2$ is the distance of point i from the origin, $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$ is the distance between points i and j, and $R = \max_i r_i$. Unless stated otherwise, the value rasterized at point i is its attribute a_i. RA, GE, and INT subsample points before constructing events, so their R is the largest radial distance among the retained points and is set to 1 if that distance is zero. The cellwise aggregation rule ϱ is selected from sum, mean, max, min, and median. PCA denotes principal component analysis.

Table S10: Point-branch Extraction operator registry.

Operator	Defining construction
Planar projection (PL)[1]	For each point $\mathbf{x}_i$ in the cloud, rasterize the projected coordinates $(\mathbf{x}_i^\top \mathbf{e}_1, \mathbf{x}_i^\top \mathbf{e}_2)$ on $[-h_u, h_u] \times [-h_v, h_v]$, where $h_u = \beta \max_i \lvert\mathbf{x}_i^\top \mathbf{e}_1\rvert$ and $h_v = \beta \max_i \lvert\mathbf{x}_i^\top \mathbf{e}_2\rvert$.
Local PCA (LP)[2]	For each point $\mathbf{x}_i$ in the cloud, take $\mathcal{N}_i = \{j : d_{ij} \le \vartheta\}$ with weights $w_{ij} = \exp(-\delta d_{ij}^2/\gamma^2)$. At each center, rasterize $\mathbf{E}_i^\top (\mathbf{x}_j - \bar{\mathbf{x}}_i)$ on $[-\beta\vartheta, \beta\vartheta]^2$, and combine centers by ϱ cellwise.
Smoothed local PCA (MLS)[2]	For each point $\mathbf{x}_i$ in the cloud, take $\mathcal{N}_i = \{j : d_{ij} \le \gamma\}$ with weights $w_{ij} = \exp(-\delta d_{ij}^2/\vartheta^2)$. At each center, rasterize $\mathbf{E}_i^\top (\mathbf{x}_j - \bar{\mathbf{x}}_i)$ on $[-\beta\gamma, \beta\gamma]^2$, and combine centers by ϱ cellwise.
Diffusion map (DM)[3] (Coifman & Lafon, 2006)	For the point cloud (X, a), set $A_{ii} = 0$ and $A_{ij} = \exp\{-[d_{ij}^2 + \delta(a_i - a_j)^2]/\vartheta^2\}$. With $\Delta = \mathrm{diag}(A\mathbf{1})$, take the eigenpairs $(\eta_k, \mathbf{q}_k)$ of $\Delta^{-1/2} A \Delta^{-1/2}$ in decreasing order and form $\mathbf{y}_k = \max(\eta_k, 0)^\beta \Delta^{-1/2} \mathbf{q}_k$ for $k = 2, 3$. Rasterize $(\mathrm{cs}_\gamma(\mathbf{y}_2; \mathbf{c}_x), \mathrm{cs}_\gamma(\mathbf{y}_3; \mathbf{c}_y))$ over its componentwise range.
Isomap (ISO)[3] (Tenenbaum et al., 2000)	Construct a graph on the point cloud by joining i and j when $d_{ij} \le \vartheta$ and assigning the edge length $\widetilde{d}_{ij} \exp(-\widetilde{d}_{ij}^2/\beta^2)$, where $\widetilde{d}_{ij}^2 = d_{ij}^2 + \delta(a_i - a_j)^2$. Apply classical multidimensional scaling (MDS) in two dimensions with ridge γI to shortest-path distances on the largest component. For its two coordinate vectors $(\mathbf{u}, \mathbf{v})$, rasterize $(\mathrm{cs}_{0.05}(\mathbf{u}; \mathbf{c}_x), \mathrm{cs}_{0.05}(\mathbf{v}; \mathbf{c}_y))$ over its componentwise range.
Locally linear embedding (LLE)[3] (Roweis & Saul, 2000)	For each point $\mathbf{x}_i$ in the cloud, take neighbors satisfying $d_{ij}^2 + \delta(a_i - a_j)^2 \le \vartheta^2$, falling back to the β nearest points when that set is empty, and run standard LLE with local covariance ridge $\gamma\ \mathrm{tr}(UU^\top) I$, where U collects the neighbor offsets $\mathbf{x}_j - \mathbf{x}_i$. For the two nonconstant embedding coordinate vectors $(\mathbf{u}, \mathbf{v})$, rasterize $(\mathrm{cs}_{0.05}(\mathbf{u}; \mathbf{1}_N), \mathrm{cs}_{0.05}(\mathbf{v}; \mathbf{1}_N))$ over its componentwise range.
Spherical projection (SPH)	Express each point of the cloud in the frame $\mathbf{e}_1 = (\cos\delta\cos\vartheta, \cos\delta\sin\vartheta, -\sin\delta)$, $\mathbf{e}_2 = (-\sin\vartheta, \cos\vartheta, 0)$, $\mathbf{e}_3 = (\sin\delta\cos\vartheta, \sin\delta\sin\vartheta, \cos\delta)$, and write $x_i' = \mathbf{x}_i^\top \mathbf{e}_1$, $y_i' = \mathbf{x}_i^\top \mathbf{e}_2$, and $z_i' = \mathbf{x}_i^\top \mathbf{e}_3$. Retain points with $\lvert r_i - \gamma\rvert \le \beta$ and rasterize $(\mathrm{atan2}(y_i', x_i'), \arccos(z_i'/r_i))$ on $[-\pi, \pi] \times [0, \pi]$.
Cylindrical projection (CYL)	In the SPH frame, retain points from the cloud satisfying $\lvert r_i^\perp - \gamma\rvert \le \beta$, where $r_i^\perp = [(x_i')^2 + (y_i')^2]^{1/2}$. Rasterize $(\mathrm{atan2}(y_i', x_i'), z_i')$ on $[-\pi, \pi]$ against the empirical range of z_i'.
Radial–angular density (RA)	Retain at most 128 radius-stratified points from the cloud (X, a). From these points, form the row histogram $\{(\min\{r_i/\vartheta, 1\}, a_i)\}$ of width β and the two-dimensional histogram $\{(\min\{(r_i + r_j)/(2\vartheta), 1\}, \phi_{ij}/\pi, (a_i + a_j)/2)\}_{i<j}$ of widths (β, γ), where ϕ_{ij} is the clipped angle between $\mathbf{x}_i$ and $\mathbf{x}_j$.

Table S10 (continued)

Operator	Defining construction
Geometry-event histogram (GE)[4]	Retain at most 64 points from the cloud (X, a) and 20,000 centered triplets formed from them. Form pair and centered-triplet histograms. Each pair's row coordinate is $(1-\vartheta)\,\mathrm{mm}(d_{ij}) + \vartheta\,\mathrm{mm}(q_{ij})$, where $q_{ij} = 1/\max\{d_{ij}, \epsilon\}$. Its weight is $(a_i + a_j)e^{-\gamma d_{ij}/d_{\max}}/2$. Triplets use $(r_{i;jk}, \phi_{jik}/\pi)$, where $r_{i;jk} = \min\{(d_{ij} + d_{ik})/(2R), 1\}$, and weight $(a_i + a_j + a_k)e^{-\gamma r_{i;jk}/r_{\max}}/3$, where $d_{\max}$ and $r_{\max}$ are the largest retained pair distance and triplet radius. Both histogram widths equal β.
Weighted interaction (INT)[4]	Retain at most 64 points from the cloud (X, a) and 20,000 centered triplets formed from them. Form separate histograms of width β for pair and centered-triplet events. A pair at distance d_{ij} has coordinate $\min\{\max\{d_{ij}, \epsilon_d\}/R, 1\}$ and weight $a_i a_j/\max\{d_{ij}, \epsilon_d\}^{\vartheta}$. A triplet has coordinates $(p_{ijk}, \phi_{jik}/\pi)$, where $p_{ijk} = \min\{\max\{(d_{ij} + d_{ik} + d_{jk})/3, \epsilon_d\}/R, 1\}$. Its weight is $a_i a_j a_k[(1-\gamma)+\gamma(1+3\cos\phi_i\cos\phi_j\cos\phi_k)]/\max\{d_{ij}d_{ik}d_{jk}, \epsilon_d^3\}^{\vartheta/3}$, where ϕ_i, ϕ_j, ϕ_k are the three internal angles.
Pairwise metric (PM)[5]	Construct G^{PM} from the point cloud (X, a) and rank-rasterize it using the stable order $\mathbf{o}$ induced by a score that interpolates, with weight γ, between normalized radius and normalized total interaction magnitude.

[1] $\mathbf{n} = (\cos\vartheta\cos\gamma, \sin\vartheta\cos\gamma, \sin\gamma)$, $\mathbf{p} = (-\sin\vartheta, \cos\vartheta, 0)$, $\mathbf{p}_\perp = \mathbf{n} \times \mathbf{p}$, $\mathbf{e}_1 = \cos\delta\,\mathbf{p} + \sin\delta\,\mathbf{p}_\perp$, $\mathbf{e}_2 = -\sin\delta\,\mathbf{p} + \cos\delta\,\mathbf{p}_\perp$.
[2] $\widehat{w}_{ij} = w_{ij}/\sum_{p\in\mathcal{N}_i} w_{ip}$, $\bar{\mathbf{x}}_i = \sum_{j\in\mathcal{N}_i} \widehat{w}_{ij}\mathbf{x}_j$, $\mathbf{E}_i = \mathrm{eig}_2^{\downarrow}(\sum_j \widehat{w}_{ij}(\mathbf{x}_j - \bar{\mathbf{x}}_i)(\mathbf{x}_j - \bar{\mathbf{x}}_i)^\top)$, where $\mathrm{eig}_2^{\downarrow}$ returns the two eigenvectors with largest eigenvalues.
[3] For the current N points, let $\mathbf{c}_x = (x_{i,1})_{i=1}^N$, $\mathbf{c}_y = (x_{i,2})_{i=1}^N$, and $\mathbf{1}_N = (1)_{i=1}^N$; for ISO these refer to its retained largest component. $\mathrm{cs}_p(\mathbf{z}; \mathbf{c})$ clips $\mathbf{z}$ to its empirical p–$(1-p)$ quantiles, then reverses the clipped vector if its dot product with $\mathbf{c}$ is negative.
[4] Min–max scaling is $\mathrm{mm}(t_e) = (t_e - \min_f t_f)/\max(\max_f t_f - \min_f t_f, \epsilon)$ over retained pair events, with $\epsilon = 10^{-12}$. For INT, $\epsilon_d = \max(10^{-6}R, 10^{-6})$. ϕ_{jik} is the clipped angle between $\mathbf{x}_j - \mathbf{x}_i$ and $\mathbf{x}_k - \mathbf{x}_i$.
[5] $G_{ij}^{\mathrm{PM}} = [(1-\delta)a_i a_j + \delta|a_i a_j|]/d_{ij}^{\vartheta}$ for $i \neq j$ and $\beta[(1-\delta)a_i + \delta|a_i|]$ for $i = j$. The ordering score is $s_i = (1-\gamma)r_i/R + \gamma\sum_j |G_{ij}^{\mathrm{PM}}|/\max_p \sum_j |G_{pj}^{\mathrm{PM}}|$ with increasing stable order $\mathbf{o}$, and $G_{o_p o_q}^{\mathrm{PM}}$ is assigned to the (p, q) rank bins of an $H \times W$ grid, merging ranks by ϱ_{PM} when the point count exceeds a grid dimension.

PL, LP, MLS, DM, ISO, LLE, SPH, CYL, and PM use hard assignment, which maps each value to a single grid cell. For coordinate-value samples $\{(u_n, v_n, w_n)\}$ and a rasterization range $\mathcal{R}$, hard assignment is

$$\left[\mathcal{G}_{\varrho}^{H,W}(\{(u_n, v_n, w_n)\}; \mathcal{R})\right]_{hk} = \varrho\{w_n : \mathrm{bin}_{\mathcal{R}}^{H,W}(u_n, v_n) = (h, k)\}.$$

Here $\mathrm{bin}_{\mathcal{R}}^{H,W}$ returns the indices of the equal-width grid cell containing (u_n, v_n). Empty cells are set to zero, and the u and v coordinates run along columns and rows, respectively. Degenerate axes are assigned to their central bins.

RA, GE, and INT use Gaussian soft assignment, which distributes event values over grid cells. The normalization rule ν is selected from total and event mean. For an event set $\mathcal{E} = \{(r_e, c_e, w_e)\}_{e=1}^{N_\mathcal{E}} \subset [0,1]^2 \times \mathbb{R}$, soft assignment is defined by

$$[\mathrm{Hist}_{\nu;\sigma_r,\sigma_c}^{H,W}(\mathcal{E})]_{hk} = \frac{\sum_e w_e \exp[-(r_e - \bar{r}_h)^2/(2\sigma_r^2) - (c_e - \bar{c}_k)^2/(2\sigma_c^2)]}{\mathrm{Norm}_\nu(N_\mathcal{E})},$$

where σ_r, σ_c are the row and column Gaussian SDs, $\bar{r}_h = (h - 1/2)/H$, $\bar{c}_k = (k - 1/2)/W$, $\mathrm{Norm}_{\mathrm{total}} = 1$, and $\mathrm{Norm}_{\mathrm{event\ mean}} = \max(N_\mathcal{E}, 1)$. $\mathrm{RowHist}_{\nu;\sigma}^{H,W}$ drops the c coordinate and replicates the result uniformly over the W columns, including a factor $1/W$. Empty event sets return zero.

For RA, GE, and INT, the two histograms I_1 and I_2, ordered as in the table, are normalized separately and combined as $I_{\mathrm{pt}} = (1-\delta)I_1 + \delta I_2$.

H.4.4 Point-branch Compression (ψ_{pt})

The Compression operator ψ_{pt} turns the matrix $M = I_{\mathrm{pt}} \in \mathbb{R}^{H\times W}$ into the descriptor $D \in \mathbb{R}^K$. Table S11 defines the eleven operator-specific mappings. The aggregation rule ζ is selected from sum, mean, max, min, and median. For cell positions, write $\mathbf{c}_{hk} = ((k - 1/2)/W, (h - 1/2)/H)$ for the normalized cell centers and $\widetilde{\mathbf{c}}_{hk} = \mathbf{c}_{hk} - (1/2, 1/2)$.

Table S11: Point-branch Compression operator registry.

Operator	Defining construction
Directional profile (DP)[1]	Profile the values M_{hk} by their directional coordinates $t_{hk} = \widetilde{\mathbf{c}}_{hk}^{\top}(\cos\lambda, \sin\lambda)$, setting $D = \mathrm{Prof}_{\varsigma}^{24}(\{(t_{hk}, M_{hk})\}_{h,k}; [-\varsigma, \varsigma])$.
Band selection (BS)[1]	Retain the cells of M with $\lvert c_{hk,1} - \lambda\rvert \le \varsigma/2$ and profile their values by $c_{hk,2}$, setting $D = \mathrm{Prof}_{\varsigma}^{24}(\{(c_{hk,2}, M_{hk}) : \lvert c_{hk,1} - \lambda\rvert \le \varsigma/2\}; [0,1])$.
Low-frequency compression (LF)[2]	Let B be the result of applying a normalized separable Gaussian blur to M with SD $\lambda\varsigma$, kernel radius $\max(1, \lfloor 3\lambda\varsigma \rfloor)$, and replicated boundaries. Set $D = \mathrm{Red}_{\varsigma}^{\mathrm{row}}(B)$, with $B = M$ when the bandwidth is zero.
Geodesic path (GD)	On the grid of M, find the minimum-cost four-connected path from $(1, \lambda)$ to (H, W) with vertex cost $1 + \varsigma\lVert\nabla M\rVert$. The descriptor D contains the values of M at 24 equally spaced arc-length positions, using the preceding path vertex.
PCA profile (PC)[1]	Compute the unweighted covariance of all $\widetilde{\mathbf{c}}_{hk}$ and let $\mathbf{e}_1, \mathbf{e}_2$ be its eigenvectors in decreasing order. For $t_{hk} = \widetilde{\mathbf{c}}_{hk}^{\top}(\cos\lambda\,\mathbf{e}_1 + \sin\lambda\,\mathbf{e}_2)$, set $D = \mathrm{Prof}_{\varsigma}^{24}(\{(t_{hk}, M_{hk})\}_{h,k}; [-\varsigma, \varsigma])$.
Polar compression (POL)[1]	With the fixed center $\mathbf{c}_0 = (1/2, 1/2)$, assign each value M_{hk} the coordinate $t_{hk} = (1-\lambda)\lVert\mathbf{c}_{hk} - \mathbf{c}_0\rVert_2 + \lambda[\mathrm{atan2}(c_{hk,2} - c_{0,2}, c_{hk,1} - c_{0,1}) + \pi]/(2\pi)$. Set $D = \mathrm{Prof}_{\varsigma}^{24}(\{(t_{hk}, M_{hk})\}_{h,k}; [0,1])$.
Row pooling (RP)[2]	Weight column k of M by setting $B_{hk} = M_{hk}\exp[-\varsigma\lvert(k-1) - \lambda\rvert/(W-1)]$, and let $D = \mathrm{Red}_{\varsigma}^{\mathrm{row}}(B)$. Use $B = M$ when $\varsigma \le 10^{-6}$.
Column pooling (CP)[2]	Weight row h of M by setting $B_{hk} = M_{hk}\exp[-\varsigma\lvert(h-1) - \lambda\rvert/(H-1)]$, and let $D = \mathrm{Red}_{\varsigma}^{\mathrm{col}}(B)$. Use $B = M$ when $\varsigma \le 10^{-6}$.
Matrix spectrum (MS)[3]	From M, set $B = M - \varsigma\bar{M}\mathbf{1}_{H\times W}$ and $s_\varsigma = 1 - \varsigma + \varsigma\max\{\mathrm{rms}(B), \epsilon\}$ for $\varsigma \in [0,1]$. Let $X = B/s_\varsigma$ if $s_\varsigma > \epsilon$, and $X = B$ otherwise. Form D by concatenating the eight leading transformed singular values of X, the eight largest transformed entry magnitudes of X, and its $(\sum, \mathrm{mean}, \mathrm{sd}, \min, \max, \ell_1, \ell_2, \text{signed mass balance})$ statistics.
Fourier moments (FM)[4]	Compute $\widehat{M} = \lvert\mathrm{DFT}_2(M)\rvert/\sqrt{HW}$ using the two-dimensional discrete Fourier transform (DFT). Vectorize the retained 4×4 low-frequency block after applying the radial gate $e^{-\sqrt{i^2+j^2}/(\lambda+\epsilon)}$ and the interpolation $(1-\varsigma)\widehat{M} + \varsigma\log(1+\widehat{M})$, then append $(\sum M, \sum\lvert M\rvert, \mathrm{mean}\, M, \mathrm{sd}\, M, m_u, m_v, v_u, v_v + c_{uv})$ to form D, where m_u, m_v, v_u, v_v, and c_{uv} are $\lvert M\rvert$-weighted spatial moments.
Multiscale summary (SC)[5]	For $s \in \{1,2,4,8\}$, set $L_s = \mathrm{Down}_s\, M$ and $\widetilde{L}_s = \mathrm{Up}_s\, L_s$. Form $E_s = \lvert M - \widetilde{L}_s\rvert$ and $T_s = (1-\varsigma)E_s + \varsigma\,\mathrm{Up}_s(\mathrm{Down}_s\, E_s)$. For each scale, multiply the six statistics $(\mathrm{mean}\, T_s, \mathrm{sd}\, T_s, \mathrm{rms}\, T_s, \mathrm{mean}\, L_s, \mathrm{sd}\, L_s, \mathrm{TV}\, \widetilde{L}_s)$ by s^λ and concatenate the four resulting vectors to form D.

[1] $[\mathrm{Prof}_{\varsigma}^{24}(\{(t_n, w_n)\}; [t_-, t_+])]_k = \zeta\{w_n : \mathrm{bin}_{[t_-,t_+]}^{24}(t_n) = k\}$, with equal-width bins and zeros for empty bins.

[2] For any $B \in \mathbb{R}^{H\times W}$, $[\mathrm{Red}_{\varsigma}^{\mathrm{row}}(B)]_h = \zeta\{B_{hk} : 1 \le k \le W\}$ and $[\mathrm{Red}_{\varsigma}^{\mathrm{col}}(B)]_k = \zeta\{B_{hk} : 1 \le h \le H\}$.

[3] Here $\bar{M} = (HW)^{-1}\sum_{h,k} M_{hk}$, $\mathrm{rms}(B) = [(HW)^{-1}\sum_{h,k} B_{hk}^2]^{1/2}$ is the root mean square (RMS), and $\epsilon = 10^{-12}$. Signed mass balance is $(P-N)/(P+N+\epsilon)$, where $P = \sum_{h,k}\max(X_{hk}, 0)$ and $N = \sum_{h,k}\max(-X_{hk}, 0)$. For each ranked nonnegative vector $\mathbf{v}$ (singular values or entry magnitudes), transform v_i to $(v_i/\max_j v_j)^\lambda$, using zeros if $\max_j v_j \le \epsilon$. Each transformed block takes its first eight values and is padded with zeros if shorter.

[4] With $\widetilde{u}_k = -1 + 2(k-1)/(W-1)$, $\widetilde{v}_h = -1 + 2(h-1)/(H-1)$, and $\mathrm{Avg}_{\lvert M\rvert}(Y) = \sum_{h,k}\lvert M_{hk}\rvert Y_{hk}/\sum_{h,k}\lvert M_{hk}\rvert$, the moments are $m_u = \mathrm{Avg}_{\lvert M\rvert}(\widetilde{u})$, $m_v = \mathrm{Avg}_{\lvert M\rvert}(\widetilde{v})$, $v_u = \mathrm{Avg}_{\lvert M\rvert}((\widetilde{u} - m_u)^2)$, $v_v = \mathrm{Avg}_{\lvert M\rvert}((\widetilde{v} - m_v)^2)$, and $c_{uv} = \mathrm{Avg}_{\lvert M\rvert}((\widetilde{u} - m_u)(\widetilde{v} - m_v))$. All five are zero when the denominator is at most ϵ.

[5] Down_s averages each block of a $\lceil H/s\rceil \times \lceil W/s\rceil$ partition, with contiguous blocks whose lengths along each axis differ by at most one. Up_s restores the original dimensions by nearest-neighbor replication. For an array A, its total variation (TV) is $\mathrm{TV}(A) = \mathrm{mean}\lvert\Delta_{\mathrm{row}} A\rvert + \mathrm{mean}\lvert\Delta_{\mathrm{col}} A\rvert$, where Δ denotes adjacent differences and an axis of length one contributes zero.

H.4.5 Address decoding and activation conditions

For the point branch, z_1 selects the Formation operator $\kappa_{\mathrm{pt},\ell}$ with $\ell = 1 + \min\{\lfloor 13 z_1 \rfloor, 12\}$, following the source order in Table S8. The coordinate z_2 sets its support parameter to $u_\kappa = (\mathrm{dir}, \mathrm{surf}, \mathrm{vol})_{1+\min\{\lfloor 3z_2\rfloor, 2\}}$, which is fixed to dir for $\ell \ge 9$.

For the Extraction stage, $1 + \min\{\lfloor 12 z_3 \rfloor, 11\}$ selects the operator of Table S10, and Table S12 decodes its parameters from $z_4, \ldots, z_7$; there, $Q_y(p)$ denotes the pth percentile of a coordinate vector y, and $r_{i,0}^{\perp} = \sqrt{x_{i1}^2 + x_{i2}^2}$ and $r_{\max,0}^{\perp} = \max_i r_{i,0}^{\perp}$ are the untilted cylindrical radii.

Table S12: Address decoding for point-branch Extraction operators.

Operator	z_4	z_5	z_6	z_7
PL	$\vartheta = 2\pi z_4$	$\beta = 0.8 + 1.2 z_5$	$\gamma = \pi z_6/4$	$\delta = \pi z_7$
LP	$\vartheta = 0.01 L_X\, 30^{z_4}$	$\beta = 0.3 + 2.7 z_5$	$\gamma = (0.2 + 0.8 z_6)\vartheta$	$\delta = 3 z_7$

Table S12 (continued)

Operator	z_4	z_5	z_6	z_7
MLS	$\vartheta = 0.01 L_X \, 20^{z_4}$	$\beta = 0.3 + 2.7 z_5$	$\gamma = (1.5 + 3.5 z_6)\vartheta$	$\delta = 0.2 + 1.3 z_7$
DM	$\vartheta = 0.01 L_X \, 30^{z_4}$	$\beta = \max\{1, 0.5 + 2.5 z_5\}$	$\gamma = 0.01 + 0.14 z_6$	$\delta = z_7$
ISO	$\vartheta = 0.01 L_X \, 30^{z_4}$	$\beta = (0.3 + 2.7 z_5)\vartheta$	$\gamma = 0.1 z_6$	$\delta = z_7$
LLE	$\vartheta = 0.01 L_X \, 30^{z_4}$	$\beta = \lfloor 2 + 8 z_5 \rfloor$	$\gamma = 10^{-5+4 z_6}$	$\delta = z_7$
SPH	$\vartheta = 2\pi z_4$	$\beta = (0.05 + 1.95 z_5) R$	$\gamma = Q_r(10 + 80 z_6)$	$\delta = \pi z_7 / 3$
CYL	$\vartheta = 2\pi z_4$	$\beta = (0.05 + 1.95 z_5) r^{\perp}_{\max,0}$	$\gamma = Q_{r^{\perp}_0}(10 + 80 z_6)$	$\delta = \pi z_7 / 3$
RA	$\vartheta = (0.8 + 1.7 z_4) R$	$\beta = 0.03 + 0.27 z_5$	$\gamma = 0.03 + 0.32 z_6$	$\delta = z_7$
GE	$\vartheta = z_4$	$\beta = 0.02 + 0.18 z_5$	$\gamma = 0.1 + 6 z_6$	$\delta = z_7$
INT	$\vartheta = 1 + 7 z_4$	$\beta = 0.02 + 0.18 z_5$	$\gamma = z_6$	$\delta = z_7$
PM	$\vartheta = 0.5 + 7.5 z_4$	$\beta = z_5$	$\gamma = z_6$	$\delta = z_7$

The common matrix dimensions are $H = \mathrm{round}(4 + 12 z_8)$ and $W = \mathrm{round}(4 + 12 z_9)$. Address coordinate z_{10} selects the entry at position $1 + \min\{\lfloor 5 z_{10} \rfloor, 4\}$ of the ordered menu $(\mathrm{sum}, \mathrm{mean}, \max, \min, \mathrm{median})$. The first eight operators and PM use that entry as the cellwise aggregation rule ϱ. RA, GE, and INT instead decode the same menu to the normalization choice $\nu = \mathrm{total}$ at positions 1, 3, and 4 and $\nu = \mathrm{event\ mean}$ at positions 2 and 5. The five menu labels therefore induce only two distinct normalization policies for those three operators.

For the Compression stage, $1 + \min\{\lfloor 11 z_{11} \rfloor, 10\}$ selects the Compression operator of Table S11, and Table S13 decodes its parameters from z_{12} and z_{13}. Where the Compression operator admits an aggregation rule, ζ is the entry at position $1 + \min\{\lfloor 5 z_{14} \rfloor, 4\}$ of $(\mathrm{sum}, \mathrm{mean}, \max, \min, \mathrm{median})$.

Table S13: Address decoding for point Compression operators.

Operator	z_{12}	z_{13}	z_{14}
DP	$\lambda = 2\pi z_{12}$	$\varsigma = 0.5 + 1.5 z_{13}$	ζ
BS	$\lambda = z_{12}$	$\varsigma = 0.02 + 0.28 z_{13}$	ζ
LF	$\lambda = z_{12}$	$\varsigma = 0.5 + 2.5 z_{13}$	ζ
GD	$\lambda = 1 + \mathrm{round}[z_{12}(W-1)]$	$\varsigma = 0.1 + 2.9 z_{13}$	–
PC	$\lambda = \pi z_{12}$	$\varsigma = 0.5 + 1.5 z_{13}$	ζ
POL	$\lambda = z_{12}$	–	ζ
RP	$\lambda = z_{12}(W-1)$	$\varsigma = 3 z_{13}$	ζ
CP	$\lambda = z_{12}(H-1)$	$\varsigma = 3 z_{13}$	ζ
MS	$\lambda = 0.35 + 2.65 z_{12}$	$\varsigma = z_{13}$	–
FM	$\lambda = 0.75 + 3.25 z_{12}$	$\varsigma = z_{13}$	–
SC	$\lambda = 2 z_{12} - 1$	$\varsigma = z_{13}$	–

H.5 Field branch

H.5.1 Branch overview

A field-branch pipeline takes the following form:

$$\mathcal{S}(\mathcal{M}) \xrightarrow{\kappa_{\mathrm{fld}}} C_{\mathrm{fld}} = (\mathbb{R}^3, f; (X, a)) \xrightarrow{\tau_{\mathrm{fld}}} I_{\mathrm{fld}} \xrightarrow{\psi_{\mathrm{fld}}} D \in \mathbb{R}^K,$$

where $f : \mathbb{R}^3 \to \mathbb{R}$ is a scalar field evaluable at arbitrary spatial coordinates, and the generating cloud (X, a) is retained to set the finite sampling domain. Unlike the other three branches, τ_{fld} has no single output shape: an Extraction operator returns a two-dimensional map in $\mathbb{R}^{H \times W}$, a voxel array in $\mathbb{R}^{N_v \times H \times W}$, a fixed-length statistics vector in $\mathbb{R}^{K_v}$, or a bounded set of locations with signed field values in $\mathbb{R}^3 \times \mathbb{R}$. The shape determines which Compression operators are admissible.

H.5.2 Field-branch Formation (κ_{fld})

For each source $s_\ell = (X_\ell, a_\ell)$ in the shared library, we define a corresponding Formation operator κ_{fld} that selects this source and constructs a scalar field on $\mathbb{R}^3$ representing its physical signal. Its

field construction q_κ is selected from the six radial kernels in Table S14 and the source-adaptive response (SAR). Its scale parameter σ sets the spatial scale of the selected construction.

For the six radial constructions in Table S14, each source point $\mathbf{x}_{\ell,i}$ contributes a radial profile centered at its position and weighted by its attribute $a_{\ell,i}$. The resulting field is $f^{(\ell)}_{q_\kappa}(\mathbf{x};\sigma) = \sum_i a_{\ell,i} k_{q_\kappa}(r_i;\sigma)$, where k_{q_κ} is the selected radial kernel and $r_i = \|\mathbf{x} - \mathbf{x}_{\ell,i}\|_2$. Here $[x]_+ = \max(x, 0)$ denotes the positive part.

Table S14: Field-kernel registry.

Kernel	$k_{q_\kappa}(r;\sigma)$
Gaussian field density (GFD)	$\exp[-r^2/(2\sigma^2)]$
Compact-support field (CSF)	$[1 - r/\sigma]^4_+$
Exponential field decay (EFD)	$\exp(-r/\sigma)$
Lorentzian field decay (LFD)	$(1 + r^2/\sigma^2)^{-1}$
Hard-sphere field (HSF)	$\mathbb{I}[r < \sigma]$
Regularized Coulomb field (RCF)	$(r^2 + \sigma^2)^{-1/2}$

For SAR, a source-dependent response replaces the shared radial profile:

$$f^{(\ell)}_{\mathrm{SAR}}(\mathbf{x};\sigma) = \begin{cases} \sum_i (a^{\mathrm{st}}_{1,i})^{1/3} / \{1 + \exp[(r_i - r^{\mathrm{st}}_{1,i})/\sigma]\}, & \ell = 1, \\ \sum_i \widetilde{a}_{\ell,i} e^{-r_i/\sigma}, & 2 \le \ell \le 7, \\ \varepsilon_{\mathrm{diel}}(\sigma)^{-1} \sum_i a_{8,i} e^{-r_i/(4\sigma)} / \sqrt{r_i^2 + \sigma^2}, & \ell = 8, \\ \sum_i a^*_{\ell,i} e^{-r_i^2/(2\sigma^2)}, & 9 \le \ell \le 13, \end{cases}$$

where $\mathrm{clip}(x, l, u)$ clips x to $[l, u]$, $a^{\mathrm{st}}_{1,i} = \mathrm{clip}(a_{1,i}, 1, 100)$, and $r^{\mathrm{st}}_{1,i} = 0.25 + 0.03(a^{\mathrm{st}}_{1,i})^{1/3}$. With $\bar{a}_\ell$ the mean of a_ℓ, set $\widetilde{a}_{\ell,i} = (a_{\ell,i} - \bar{a}_\ell)/\,\mathrm{sd}(a_\ell)$ when the cloud has at least two values and $\mathrm{sd}(a_\ell) > \epsilon$; otherwise $\widetilde{a}_{\ell,i} = a_{\ell,i}$. Set $\varepsilon_{\mathrm{diel}}(\sigma) = 1 + 3\sigma/(\sigma + 1)$; for $\ell \ge 9$, $a^*_{\ell,i} = a_{\ell,i}$ unless the source is numerically zero, in which case $a^*_{\ell,i} = 1$.

H.5.3 Field-branch Extraction (τ_{fld})

The Extraction operator τ_{fld} maps the carrier $C_{\mathrm{fld}} = (\mathbb{R}^3, f; (X, a))$ to the intermediate I_{fld}. Table S15 defines seven operators that sample, integrate, or summarize f over spatial domains determined by the source geometry X and the operator parameters. Here ϑ scales the sampling domain, while β is the slice angle for FS or the inner-radius fraction for SI, CI, and RS.

Write $R_X = \max_i \|\mathbf{x}_i\|_2$, $R^{\parallel}_X = \max_i |x_{i3}|$, and $R^{\perp}_X = \max_i (x^2_{i1} + x^2_{i2})^{1/2}$ for the radial, axial, and transverse extents of X. Its scaled bounding box is $\mathrm{Box}_\vartheta(X) = [\mathbf{c}_X - \vartheta \mathbf{h}_X, \mathbf{c}_X + \vartheta \mathbf{h}_X]$, where $\mathbf{c}_X$ and $\mathbf{h}_X$ are the midpoint and half-separation of the coordinatewise extrema of X, the latter raised componentwise to at least $1/2$. Let $\mathcal{L}_n[t_0, t_1]$ denote n equally spaced samples of $[t_0, t_1]$ with the upper endpoint omitted.

Table S15: Field-branch Extraction operator registry.

Operator	Defining construction	Intermediate
Planar slice (FS)	Sample $f(u_k \mathbf{e}_1 + v_h \mathbf{e}_2)$ for $u_k \in \mathcal{L}_W[-h_1, h_1]$ and $v_h \in \mathcal{L}_H[-h_2, h_2]$, where $\mathbf{e}_1 = (\cos\beta, 0, -\sin\beta)$, $\mathbf{e}_2 = (0, 1, 0)$, and $h_1 = \vartheta \max_i \lvert\mathbf{x}_i^\top \mathbf{e}_1\rvert$ and $h_2 = \vartheta \max_i \lvert\mathbf{x}_i^\top \mathbf{e}_2\rvert$.	$H \times W$ map
Spherical integration (SI)	Integrate $f(r\mathbf{d}(\xi_k, \phi_h))r^2$ over $r \in [\beta\vartheta R_X, \vartheta R_X]$ by 64-point trapezoidal quadrature, for $\xi_k \in \mathcal{L}_W[-\pi, \pi]$ and $\phi_h \in \mathcal{L}_H[0, \pi]$, where $\mathbf{d} = (\sin\phi\cos\xi, \sin\phi\sin\xi, \cos\phi)$.	$H \times W$ map
Cylindrical integration (CI)	Integrate $f(r_\perp \cos\xi_k, r_\perp \sin\xi_k, z_h) r_\perp$ over $r_\perp \in [\beta\vartheta R^{\perp}_X, \vartheta R^{\perp}_X]$ by the same quadrature, for $z_h \in \mathcal{L}_H[-\vartheta R^{\parallel}_X, \vartheta R^{\parallel}_X]$ and $\xi_k \in \mathcal{L}_W[-\pi, \pi]$.	$H \times W$ map
Radial-shell map (RS)	Average $f(r_h \mathbf{d}(2\pi n/32, \phi_k))$ over $n = 0, \ldots, 31$, for $r_h \in \mathcal{L}_H[\beta\vartheta R_X, \vartheta R_X]$ and $\phi_k \in \mathcal{L}_W[0, \pi]$.	$H \times W$ map
Voxel grid (VX)	Sample f on the $N_v \times H \times W$ lattice spanning $\mathrm{Box}_\vartheta(X)$, with upper faces omitted.	voxel array

Table S15 (continued)

Operator	Defining construction	Intermediate
Global statistics (ST)[1]	Apply the 21-component voxel summary $\mathbf{s}_3$ to f sampled on the $N_s \times H_s \times W_s$ lattice over $\mathrm{Box}_\vartheta(X)$, then $\mathcal{P}_{K_v}$.	K_v vector
Extreme nodes (EX)	Retain the at most $N_{\max}$ sites of largest $\lvert f\rvert$ on the $N_e \times H_e \times W_e$ lattice over $\mathrm{Box}_\vartheta(X)$, together with their signed values. An all-zero field gives the empty set.	point set

[1] $\mathcal{P}_K$ replaces nonfinite entries by zero and pads or truncates to length K. For a voxel array V, $\mathbf{s}_3$ contains, in order, $(\sum, \mathrm{mean}, \mathrm{sd}, \min, \max, \ell_1, \ell_2)$ of the voxel values; positive and negative masses $P = \sum[V]_+$, $N = \sum[-V]_+$ and signed mass balance $(P - N)/(P + N + \epsilon)$; quantiles at $p = (0.05, 0.25, 0.50, 0.75, 0.95)$; and the $\lvert V\rvert$-weighted centroid (c_x, c_y, c_z) and coordinatewise variances (v_x, v_y, v_z) on $[-1, 1]^3$. Zero mass gives six zero spatial moments; $\mathcal{P}_{24}$ appends three zeros to the 21 components.

For degenerate extents, the FS half-widths and the CI axial half-width and transverse outer radius are each bounded below by 10^{-6}. SI and RS use outer radius 5 when $R_X = 0$. In SI, CI, and RS, β multiplies the resulting outer radius to set the lower radial limit.

H.5.4 Field-branch Compression (ψ_{fld})

The Compression operator ψ_{fld} maps the intermediate I_{fld} to the descriptor $D \in \mathbb{R}^K$. The type of I_{fld} determines the admissible Compression operators. Matrix intermediates use one of the eleven point-branch operators in Table S11. Where applicable, the aggregation rule ζ is selected from sum, mean, max, min, and median. For voxel arrays, vectors, and point sets, Table S16 defines six operators, each producing a 24-component descriptor.

In Table S16, V denotes a voxel array, $\mathbf{v}$ a vector, and $\{(\mathbf{x}_i^E, a_i^E)\}$ a point set with locations $\mathbf{x}_i^E$ and signed values a_i^E.

Table S16: Non-matrix field Compression operator registry.

Operator	Defining construction	Input
Voxel moments (VM)[1]	Summarize the voxel array V with $\mathbf{s}_3$ and apply the finalizer to obtain $D = \mathcal{P}_{24}(\mathbf{s}_3(V))$.	voxel array
Voxel spectrum (VS)[1]	For the voxel array V, set $V_\varsigma = V - \varsigma\bar{V}$ with $\bar{V}$ the mean of V, and $\widehat{V}_\varsigma = \lvert\,\mathrm{DFT}_3(V_\varsigma)\rvert/\sqrt{N_v HW}$ using the three-dimensional DFT. Let $m = \max \widehat{V}_\varsigma$ and transform each spectral entry v to $(v/m)^\lambda$ if $m > \epsilon$, or leave v unchanged otherwise. Concatenate the 16 largest transformed entries with the mean, SD, extrema, and ℓ_1/ℓ_2 norms of V_ς, and the mean and SD of $\widehat{V}_\varsigma$. Apply $\mathcal{P}_{24}$ to the concatenated vector to obtain D.	voxel array
Multiscale summary (VSC)[2]	Apply the block averaging, replication, and residual mixing of SC, with mixing weight ς, to the voxel array V in three dimensions at scales $s \in \{1, 2, 4, 8\}$. For each scale, form a six-component vector containing the mean, SD, and RMS of the mixed residual, the mean and SD of the coarse array, and the three-axis TV of the replicated array. Multiply each vector by s^λ and concatenate the four vectors to obtain D.	voxel array
Vector finalization (DF)[1]	Apply the finalizer to the vector $\mathbf{v}$ to obtain $D = \mathcal{P}_{24}(\mathbf{v})$.	K_v vector
Pair histogram (EP)	From the point set $\{(\mathbf{x}_i^E, a_i^E)\}$, construct the weighted pair-distance histogram $\{\sum_{i<j} \lvert a_i^E a_j^E\rvert \mathbb{I}[r_{ij} \in \mathcal{I}_k]\}_{k=1}^{24}$, where $\mathcal{I}_1, \ldots, \mathcal{I}_{24}$ are equal bins on $[0, \max(\varsigma r_*, \epsilon)]$, $r_{ij} = \lVert\mathbf{x}_i^E - \mathbf{x}_j^E\rVert_2$, and $r_* = \max_{i<j} r_{ij}$. Raise the histogram entries componentwise to λ to obtain D. Pairs beyond the range are omitted, and fewer than two nodes give zero.	point set
Node summary (ES)[1]	Summarize the point set $\{(\mathbf{x}_i^E, a_i^E)\}$ by concatenating, in order, the node count, the statistics $(\sum, \mathrm{mean}, \mathrm{sd}, \min, \max, \ell_1, \ell_2)$ of the signed values a_i^E, the statistics $(\mathrm{mean}, \mathrm{sd}, \min, \max)$ of the radii $\lVert\mathbf{x}_i^E\rVert_2$, the $\lvert a^E\rvert$-weighted centroid (c_x, c_y, c_z) and coordinatewise variances (v_x, v_y, v_z), and the signed-value quantiles at $p = (0.05, 0.25, 0.50, 0.75, 0.95)$. Apply $\mathcal{P}_{24}$ to pad these 23 components and obtain D. Return zero if $\sum_i \lvert a_i^E\rvert \le \epsilon$.	point set

[1] The voxel summary $\mathbf{s}_3$ and finalizer $\mathcal{P}_K$ are defined in the note to Table S15.
[2] SC is the point-branch multiscale summary defined in Table S11.

H.5.5 Address decoding and activation conditions

For the field branch, z_1 selects the Formation operator $\kappa_{\mathrm{fld},\ell}$ with $\ell = 1 + \min\{\lfloor 13 z_1 \rfloor, 12\}$, following the source order in Table S8. The coordinate z_2 sets its construction parameter to $q_\kappa = (\mathrm{GFD}, \mathrm{CSF}, \mathrm{EFD}, \mathrm{LFD}, \mathrm{HSF}, \mathrm{RCF}, \mathrm{SAR})_{1+\min\{\lfloor 7 z_2 \rfloor, 6\}}$, covering the six radial kernels in Table S14 and SAR. The coordinate z_4 sets its spatial scale to $\sigma = 0.01 \max(L_X, \epsilon) 30^{z_4} > 0$.

For the Extraction stage, $1 + \min\{\lfloor 7z_3 \rfloor, 6\}$ selects an Extraction operator from Table S15, and Table S17 decodes its parameters: every Extraction operator decodes the domain scale $\vartheta = 0.8 + 1.7z_5$, and the table gives the slice angle or inner-radius fraction β and the effective size in terms of the nominal grid dimensions $H_0 = \mathrm{round}(4 + 12z_8)$ and $W_0 = \mathrm{round}(4 + 12z_9)$.

Table S17: Address decoding for field-branch Extraction operators.

Extraction	z_6	Effective size
FS	$\beta = \pi z_6/2$	$H_0 \times W_0$
SI	$\beta = z_6$	$H_0 \times W_0$
CI	$\beta = z_6$	$H_0 \times W_0$
RS	$\beta = z_6$	$H_0 \times W_0$
VX	–	$N_v = H = H_0, W = W_0$
ST	–	$K_v = 24, N_s = H_s = \min\{H_0, 12\}, W_s = \min\{W_0, 12\}$
EX	–	$N_{\max} = 24, N_e = H_e = H_0, W_e = W_0$

The shape of the Extraction intermediate determines which Compression operators are admissible. For a matrix intermediate, the choices are the eleven point-branch operators, whose parameters are decoded as in Table S13, including the ζ pooling menu. For a voxel array, vector, or point set, the choices are the operators in Table S16, and their parameters are decoded by Table S18. Within the compatible menu in table order, z_{11} selects entry $1 + \min\{\lfloor n_\psi z_{11} \rfloor, n_\psi - 1\}$, where the menu size n_ψ is 11, 3, 1, or 2 for matrices, voxel arrays, vectors, or point sets, respectively.

Table S18: Address decoding for non-matrix field Compression operators by intermediate type.

Operator	z_{12}	z_{13}
VM	–	–
VS	$\lambda = 0.35 + 2.65z_{12}$	$\varsigma = z_{13}$
VSC	$\lambda = 2z_{12} - 1$	$\varsigma = z_{13}$
DF	–	–
EP	$\lambda = 0.5 + 2z_{12}$	$\varsigma = 0.5 + 1.5z_{13}$
ES	–	–

H.6 Expansion branch

H.6.1 Branch overview

An expansion-branch pipeline takes the following form:

$$\mathcal{S}(\mathcal{M}) \xrightarrow{\kappa_{\exp}} C_{\exp} = (X, \mathbf{W}; \varnothing) \xrightarrow{\tau_{\exp}} I_{\exp} = (B_c)_{c=1}^{N} \xrightarrow{\psi_{\exp}} D \in \mathbb{R}^K,$$

where $X \in \mathbb{R}^{N\times 3}$ holds the point coordinates, $\mathbf{W} \in \mathbb{R}^{N_q \times N}$ holds one weight per channel and point, and $B_c \in \mathbb{R}^{N_q \times N_r \times N_d \times N_k}$ is the feature tensor for center c. Its channel, radial, degree, and component axes are interpreted according to the selected Extraction operator. The channel index q starts at 1, while r, d, k start at 0. The Extraction operator $\tau_{\exp}$ encodes every point's neighborhood using radial and angular expansions, moment constructions, or correlation and spectral features.

H.6.2 Expansion-branch Formation ($\kappa_{\exp}$)

For each source $s_\ell = (X_\ell, a_\ell)$ in the shared library, we define a corresponding Formation operator $\kappa_{\exp}$ that selects this source and constructs a point cloud with channel weights. It has no additional parameters. The carrier retains the source coordinates, $X = X_\ell$, and assigns channel weights according to the selected source. For the atomic-number source ($\ell = 1$), let Z_i denote the atomic number at point i and $Z^{(1)} < \cdots < Z^{(N_q)}$ the distinct atomic numbers across the dataset. The channel weights are

$$\mathbf{W}_{qi} = \begin{cases} \mathbb{I}[Z_i = Z^{(q)}], & \ell = 1, \\ a_{\ell,i}, & 2 \le \ell \le 13, \quad q = 1. \end{cases}$$

The dataset-wide vocabulary fixes the number of indicator channels and the channel width of downstream tensors across molecules. Every other source uses one scalar channel ($N_q = 1$) containing its attributes.

H.6.3 Expansion-branch Extraction (τ_{exp})

The Extraction operator τ_{exp} maps the carrier $C_{\mathrm{exp}} = (X, \mathbf{W}; \varnothing)$ to the intermediate $I_{\mathrm{exp}} = (B_c)_{c=1}^{N}$. For each point $\mathbf{x}_c \in X$, it defines a neighborhood in coordinates relative to that point and transforms the associated channel weights. Each operator then constructs the local tensor B_c from point profiles, interaction events, or neighborhood statistics, as specified in Table S19.

For center c and point i, let $\mathbf{x}_{ci} = \mathbf{x}_i - \mathbf{x}_c$, $r_{ci} = \|\mathbf{x}_{ci}\|_2$, and $\mathbf{n}_{ci} = \mathbf{x}_{ci}/r_{ci}$ when $r_{ci} > \epsilon$, with $\mathbf{n}_{ci} = \mathbf{0}$ otherwise. For a direction $\mathbf{n} = (n_x, n_y, n_z)$, define its azimuth and polar angle by $\xi(\mathbf{n}) = \operatorname{atan2}(n_y, n_x)$ and $\phi(\mathbf{n}) = \arccos(n_z)$. At zero direction, use $\xi(\mathbf{0}) = 0$ and $\phi(\mathbf{0}) = \pi/2$. The same convention applies to pair displacements in PS. The neighborhood of center c is $\mathcal{N}_c(\beta) = \{i : r_{ci} \le \beta\}$, which includes every point when $\beta = \infty$. Let $r_c^{\max} = \max_{i \in \mathcal{N}_c(\beta)} r_{ci}$. The normalized radius is $t_{ci}^{(\vartheta)} = \mathrm{clip}(r_{ci}/(\vartheta r_c^{\max}), 0, 1)$ when $r_c^{\max} > 0$, and $t_{ci}^{(\vartheta)} = 0$ otherwise. The value transform γ is selected from raw, absolute, signed-root, signed-log, and centered. For each channel q with $w = \mathbf{W}_{qi}$, the corresponding values $\widetilde{W}_{qi}$ are w, $|w|$, $\mathrm{sgn}(w)\sqrt{|w|}$, $\mathrm{sgn}(w)\log(1 + |w|)$, and w centered by its channel mean over $\mathcal{N}_c(\beta)$, respectively.

Table S19: Expansion-branch Extraction operator registry.

Extraction	Defining construction
Radial–harmonic (RH)[1,3]	Set $[B_c]_{qrdk} = \sum_{i \in \mathcal{N}_c(\beta)} \widetilde{W}_{qi} g_r(t_{ci}^{(\vartheta)}; 0.22)\, h_{dk}(\mathbf{n}_{ci})\, e^{-\delta t_{ci}^{(\vartheta)}}$.
Volumetric orthogonal (VO)[2,3]	Set $[B_c]_{qrdk} = \sum_{i \in \mathcal{N}_c(\beta)} \widetilde{W}_{qi} o_r(t_{ci}^{(\vartheta)}; 1/2)\, h_{dk}(\mathbf{n}_{ci})/(1 + \delta r)$.
Orientational harmonic (OH)[1,3]	Set $[B_c]_{qrdk} = \sum_{i \in \mathcal{N}_c(\beta)} \widetilde{W}_{qi}[(1 - \delta) + \delta g_r(t_{ci}^{(\vartheta)}; 0.094)]\, h_{dk}(\mathbf{n}_{ci})$.
Tensor moments (TM)[1]	Set $[B_c]_{qrdk} = \sum_{i \in \mathcal{N}_c(\beta)} \widetilde{W}_{qi} g_r(t_{ci}^{(\vartheta)}; 0.22)(1 + \delta t_{ci}^{(\vartheta)})\, \mathrm{mon}_{dk}(\mathbf{n}_{ci})$, where $\mathrm{mon}_{dk}(\mathbf{n}) = n_x^{p_x} n_y^{p_y} n_z^{p_z}$ runs over the exponents with $p_x + p_y + p_z = d$ and cycles through them when N_k is larger.
Many-body distribution (MBD)[1,6,7,8]	Construct $[B_c]_{qrdk}$ by summing the capped pair events $\bar{w}_{q,ij}\, g_r(\widehat{\Delta}_{ij}; 0.20)\, \mathrm{trig}_k(\omega_{ij})/(d + 1)$ and, with multiplier δ, the capped triplet events $\bar{w}_{q,ijp}\, g_r(\widehat{\Delta}_{ijp}; 0.20)[(d+1)/N_d] \cos[(k + 1)\omega_{ijp}]$, where trig_k alternates $\cos[(k + 1)\omega]$ and $\sin[(k + 1)\omega]$, ω_{ij} is the angle between $\mathbf{x}_{ci}$ and $\mathbf{x}_{cj}$, and ω_{ijp} is the angle at i between $\mathbf{x}_{cj} - \mathbf{x}_{ci}$ and $\mathbf{x}_{cp} - \mathbf{x}_{ci}$. To evaluate the angles, we bound the denominator of each cosine by ϵ. We set $\omega_{ij} = \pi/2$ if either vector in a pair is zero, and $\omega_{ijp} = 0$ if either incident edge of a triplet has length at most ϵ. Both event types remain in the sum.
Hyperspherical-like (HS)[2]	Set $[B_c]_{qrdk} = \sum_{i \in \mathcal{N}_c(\beta)} \widetilde{W}_{qi} o_r(t_{ci}^{(\vartheta)}; \delta)\, \mathrm{hyp}_{dk}(\chi_{ci}, \xi(\mathbf{n}_{ci}), \phi(\mathbf{n}_{ci}))$, where $\chi_{ci} = \pi(t_{ci}^{(\vartheta)} + [t_{ci}^{(\vartheta)}]^2)/2$ and hyp_{dk} alternates the cosine and the sine of $(d + 1)\chi + [(k \bmod 3) + 1]\xi + [(\lfloor k/3 \rfloor \bmod 3) + 1]\phi$.
Pair-distance spectral (PS)[1,2,6,7,8]	Set $[B_c]_{qrdk}$ to the sum over capped pair events of $\bar{w}_{q,ij}\, g_r(\widehat{\Delta}_{ij}; 0.22)\, o_d(\widehat{\Delta}_{ij}; \delta)\, \mathrm{ang}_k(\widehat{\mathbf{x}}_{ij})$, where $\widehat{\mathbf{x}}_{ij}$ is the unit vector along $\mathbf{x}_{cj} - \mathbf{x}_{ci}$ and ang_k cycles through $\cos[(k+1)\xi], \sin[(k+1)\xi]$, and $\cos[(k + 1)\phi]$.
Shape and inertia (SM)[1,4,5]	Combine a shell moment and a point-averaged bank to form $[B_c]_{qrdk}$, with mixing weight δ on the latter. The shell moment is $\sum_{i \in \mathcal{N}_c(\beta)} \widetilde{W}_{qi}\, g_r(t_{ci}^{(\vartheta)}; 0.22)[t_{ci}^{(\vartheta)}]^d\, \mathrm{sh}_k(\mathbf{x}_{ci})$. The point-averaged bank is $\lvert\mathcal{N}_c(\beta)\rvert^{-1} \sum_i g_r(t_{ci}^{(\vartheta)}; 0.22)$ scaled by $(\mathrm{tr}\,\Sigma_q, (\varpi_1 - \varpi_3)/(\mathrm{tr}\,\Sigma_q + \epsilon), r_q^{\mathrm{rms}}, \varpi_1, \varpi_2, \varpi_3)_{d \bmod 6}$, where $\varpi_1 \ge \varpi_2 \ge \varpi_3$ are the eigenvalues of Σ_q and sh_k cycles through $1, n_x, n_y, n_z, n_x n_y, n_x n_z, n_y n_z, n_x^2 - n_y^2, 3n_z^2 - 1, t^{(\vartheta)}, [t^{(\vartheta)}]^2$, and the unscaled radius r_{ci}, with $\mathbf{n} = \mathbf{n}_{ci}$ and $t^{(\vartheta)} = t_{ci}^{(\vartheta)}$.
Leverage autocorrelation (LA)[1,4,6,7]	Construct $[B_c]_{qrdk}$ by summing the capped pair events $\mathrm{sgn}(\widetilde{W}_{qi} + \widetilde{W}_{qj})\sqrt{\lvert\widetilde{W}_{qi}\widetilde{W}_{qj}\rvert + \epsilon}\,[(1 - \delta)(\mathrm{lev}_{qi} + \mathrm{lev}_{qj})/2 + \delta\, \mathrm{lev}_{qi}\, \mathrm{lev}_{qj}]\, g_r(\widehat{\Delta}_{ij}; \vartheta) \cos[\pi(d+1)\widehat{\Delta}_{ij}]\, \mathrm{osc}_k(\widehat{\Delta}_{ij})$, where osc_k alternates $\cos[\pi(k + 1)\widehat{\Delta}]$ and $\sin[\pi(k + 1)\widehat{\Delta}]$, and lev_{qi} is the squared Mahalanobis leverage of point i under the $\lvert\widetilde{W}_q\rvert$-weighted neighborhood covariance with ridge $0.005001 I_3$. It is divided by the neighborhood maximum only when that maximum exceeds ϵ; otherwise the unnormalized leverage is retained.

[1] The Gaussian radial function is $g_r(t; \sigma) = \exp[-(t - (r + 1/2)/N_r)^2/(2\sigma^2)]$.

[2] The oscillatory radial function is $o_n(t; \eta) = (1 - \eta)\cos[\pi(n + 1)t] + \eta \sin[\pi(n + 1)t]$.

[3] The real angular functions use $\iota_k = k - \lfloor N_k/2 \rfloor$, $\xi = \xi(\mathbf{n})$, and the degree-d Legendre polynomial Leg_d:

$$h_{dk}(\mathbf{n}) = \begin{cases} \mathrm{Leg}_d(n_z), & \iota_k = 0, \\ (1 - n_z^2)^{\iota_k/2} \cos(\iota_k \xi), & 0 < \iota_k \le d, \\ (1 - n_z^2)^{|\iota_k|/2} \sin(|\iota_k| \xi), & -d \le \iota_k < 0, \\ 0, & |\iota_k| > d. \end{cases}$$

Table S19 (continued)

[4] The weighted neighborhood statistics use $a_{qi} = |\widetilde{W}_{qi}|$, $A_q = \sum_i a_{qi}$, and $\mathbf{m}_q = \sum_i a_{qi}\mathbf{x}_{ci}/(A_q + \epsilon)$. The weighted covariance is $\Sigma_q = \sum_i a_{qi}(\mathbf{x}_{ci} - \mathbf{m}_q)(\mathbf{x}_{ci} - \mathbf{m}_q)^\top/(A_q + \epsilon)$. All sums run over $\mathcal{N}_c(\beta)$. When $A_q = 0$, $\mathbf{m}_q = \mathbf{0}$ and $\Sigma_q = 0$, and the output is zero for that channel.

[5] The weighted RMS radius is $r_q^{\mathrm{rms}} = (\sum_{i\in\mathcal{N}_c(\beta)} a_{qi} r_{ci}^2/(A_q + \epsilon))^{1/2}$.

[6] Event thinning uses a deterministic uniform stride to retain at most 1024 pair or triplet events per center.

[7] Distance normalization caps $\widehat{\Delta}$ at one. Its divisor is the largest retained pair distance when only pairs are used and $r_c^{\max}$ when triplets are also used. The unnormalized triplet distance is the mean length of the two edges incident to i.

[8] The event weights $\bar{w}_{q,ij}$ and $\bar{w}_{q,ijp}$ are the arithmetic means of the participating $\widetilde{W}_{q\cdot}$ values.

H.6.4 Expansion-branch Compression (ψ_{exp})

The Compression operator ψ_{exp} maps the local feature tensors $I_{\mathrm{exp}} = (B_c)_{c=1}^N$ to the descriptor $D \in \mathbb{R}^K$. The selected operator in Table S20 is applied to each B_c, and the resulting local vectors $\mathbf{u}_c$ are averaged as $\bar{\mathbf{u}} = N^{-1}\sum_{c=1}^N \mathbf{u}_c$. Here vec flattens an array into a vector.

Table S20: Expansion-branch Compression operator registry.

Operator	Defining construction
Energy summary (CES)[1,3,4]	Concatenate the twelve largest entries of $\mathrm{vec}\, E^{(\lambda,\varsigma)}$ in decreasing order, $\mathbf{s}_8(E^{(\lambda,\varsigma)})$, and the first four entries of $\mathrm{vec}\, E^{(\lambda,\varsigma)}$, then apply $\mathcal{F}_\zeta$.
Gram spectrum (GS)[1,2,3]	Reshape B_c to $Y \in \mathbb{R}^{(N_q N_r N_d)\times N_k}$. Let $\bar{Y}_{\mathrm{row}}$ repeat each row mean across its N_k entries, and set $Y_\varsigma = Y - \varsigma\bar{Y}_{\mathrm{row}}$ and $G_\varsigma = Y_\varsigma Y_\varsigma^\top/N_k$. Concatenate ten leading SPow_λ eigenvalues, six SPow_λ diagonal entries, and $\mathbf{s}_8(G_\varsigma)$, then apply $\mathcal{F}_\zeta$.
Unfolding SVD (US)[1,2,3]	Unfold B_c along axes λ, $\lambda+1$, and $\lambda+2$, taken modulo four, for $\lambda \in \{0,1,2,3\}$. Compute a singular value decomposition (SVD) of each unfolding and apply SPow_ς to its six leading singular values. Concatenate these values with $\mathbf{s}_8(B_c)$ and apply $\mathcal{F}_\zeta$.
Axis moments (AM)[1,2,3]	Form $Y = (1-\varsigma)B_c + \varsigma|B_c|$. For each of its four axes, average Y along that axis and take the mean, SD, and maximum of the result, giving twelve values. Concatenate them with $\mathbf{s}_8(Y)$ and the four leading SPow_λ values of $\sum_{q,k} Y_{qrdk}^2$, then apply $\mathcal{F}_\zeta$.
Correlation sketch (CSK)[3]	Set $y = \mathrm{vec}\, B_c - \varsigma\bar{B}_c$, where $\bar{B}_c$ is the mean of the entries of B_c. Divide y by $\|y\|_2$ when this norm exceeds ϵ. Use its circular autocorrelation for $\lambda < 1/2$. Otherwise, interpolate by λ between that autocorrelation and the circular cross-correlation of y^2 with y. Retain twelve uniformly strided lags of the resulting correlation, six leading degree energies, and six leading component energies, then apply $\mathcal{F}_\zeta$.
Multiscale summary (MSP)[3]	For each $s \in \{1,2,4,8\}$, block-average B_c along the radial and degree axes. Collect the mean and SD of the coarse tensor, the RMS of B_c, a mixture of the coarse SD and the radial-plus-degree gradient with respective weights $1-\varsigma$ and ς, and the radial and degree gradients separately. Scale these six values by s^λ, concatenate the four resulting vectors, and apply $\mathcal{F}_\zeta$.
Raw tensor (RF)[5]	Let $v = \mathrm{vec}\, B_c$. Return $(1-\varsigma)v + \varsigma v/\|v\|_2$ when $\|v\|_2 > \epsilon$, and return v otherwise.
Full energy (CEF)[4,5]	Return $\mathrm{vec}\, E^{(\lambda,\varsigma)}$.
Full Gram (GF)[2,5]	Form G_ς as for GS and return the upper triangle of $\mathrm{SPow}_\lambda(G_\varsigma)$.
Channel-pair power (CPP)[2,5]	Reshape $[B_c]_{:,:,d,:}$ at each degree d to $Y_d \in \mathbb{R}^{(N_q N_r)\times N_k}$ and subtract ς times each row mean to obtain $Y_{d,\varsigma}$. Apply SPow_λ to $Y_{d,\varsigma}Y_{d,\varsigma}^\top/N_k$, retain its upper triangle, and concatenate these vectors over d.

[1] The summary $\mathbf{s}_8$ contains the sum, mean, SD, minimum, maximum, ℓ_1 norm, RMS, and signed mass balance of its input entries v_j. The last is $\sum_j v_j/(\sum_j |v_j| + \epsilon)$.

[2] The signed-power transform $\mathrm{SPow}_p(v)$ uses $M = \max_j |v_j|$ for a vector or matrix v. Its entrywise rule maps v_j to $\mathrm{sgn}(v_j)(|v_j|/M)^p$ when $M > \epsilon$. Compact readouts return zero when $M \le \epsilon$. Full Gram readouts retain v when $M \le \epsilon$ or $|p-1| \le \epsilon$, and use the entrywise rule otherwise.

[3] The finalization rule ζ is selected from identity, centered RMS, max-absolute, signed-log, and ℓ_2 normalization. The finalizer $\mathcal{F}_\zeta$ pads or truncates the vector to 24 components, replaces nonfinite entries by zero, and applies ζ.

[4] The energy array is $E_{qrd}^{(\lambda,\varsigma)} = (1-\varsigma)(N_k^{-1}\sum_k |[B_c]_{qrdk}|^\lambda + \epsilon)^{1/\lambda} + \varsigma \max_k |[B_c]_{qrdk}|$.

[5] PCA maps the center-averaged vector $\bar{\mathbf{u}}$ to D. The projection is fitted on the training split to retain 80% of the variance and reused unchanged on validation and test data.

H.6.5 Address decoding and activation conditions

For the expansion branch, z_1 selects the Formation operator $\kappa_{\mathrm{exp},\ell}$ with $\ell = 1 + \min\{\lfloor 13 z_1 \rfloor, 12\}$, following the source order in Table S8.

For the Extraction stage, $1 + \min\{\lfloor 9 z_3 \rfloor, 8\}$ selects an operator from Table S19. Table S21 gives the operator-specific decoding of ϑ and δ from z_4 and z_7. All operators use z_6 to select the value

transform $\gamma = (\text{raw}, \text{absolute}, \text{signed-root}, \text{signed-log}, \text{centered})_{1+\min\{\lfloor 5z_6 \rfloor, 4\}}$ and z_5 to set the neighborhood cutoff

$$\beta = \begin{cases} (0.10 + 0.50z_5)L_X, & N \geq 15 \text{ and } L_X > \epsilon, \\ \infty, & \text{otherwise.} \end{cases}$$

The radial and degree budgets are $N_r = \min\{\text{round}(4 + 12z_8), 10\}$ and $N_d = \min\{\text{round}(4 + 12z_9), 8\}$. The component budget follows from N_d as $N_k = \min\{\max(3, 2N_d + 1), 17\}$.

For the Compression stage, $1 + \min\{\lfloor 10z_{11} \rfloor, 9\}$ selects an operator from Table S20. Table S22 gives the decoding of its parameters from z_{12} and z_{13}. For operators that use $\mathcal{F}_\zeta$, the index $1 + \min\{\lfloor 5z_{14} \rfloor, 4\}$ selects ζ from identity, centered RMS, max-absolute, signed-log, and ℓ_2 normalization, in that order.

Table S21: Address decoding for expansion-branch Extraction operators.

Extraction	z_4	z_7
RH	$\vartheta = 0.75 + 2.25z_4$	$\delta = 2z_7$
VO	$\vartheta = 0.75 + 2.25z_4$	$\delta = z_7$
OH	$\vartheta = 0.75 + 2.25z_4$	$\delta = z_7$
TM	$\vartheta = 0.75 + 2.25z_4$	$\delta = z_7$
MBD	–	$\delta = z_7$
HS	$\vartheta = 0.75 + 2.25z_4$	$\delta = z_7$
PS	–	$\delta = z_7$
SM	$\vartheta = 0.75 + 2.25z_4$	$\delta = z_7$
LA	$\vartheta = 0.04 + 0.36z_4$	$\delta = z_7$

Table S22: Address decoding for expansion Compression operators.

Operator	z_{12}	z_{13}	z_{14}
CES	$\lambda = 1 + 3z_{12}$	$\varsigma = z_{13}$	ζ
GS	$\lambda = 0.35 + 2.65z_{12}$	$\varsigma = z_{13}$	ζ
US	$\lambda = \min\{\lfloor 4z_{12} \rfloor, 3\}$	$\varsigma = 0.5 + 2.5z_{13}$	ζ
AM	$\lambda = 0.5 + 2z_{12}$	$\varsigma = z_{13}$	ζ
CSK	$\lambda = z_{12}$	$\varsigma = z_{13}$	ζ
MSP	$\lambda = 2z_{12} - 1$	$\varsigma = z_{13}$	ζ
RF	–	$\varsigma = z_{13}$	–
CEF	$\lambda = 1 + 3z_{12}$	$\varsigma = z_{13}$	–
GF	$\lambda = 1 + 2z_{12}$	$\varsigma = z_{13}$	–
CPP	$\lambda = 1 + 2z_{12}$	$\varsigma = z_{13}$	–

H.7 STATISTICS BRANCH

H.7.1 BRANCH OVERVIEW

A statistics-branch pipeline takes the following form:

$$\begin{aligned} \mathcal{S}(\mathcal{M}) &\xrightarrow{\kappa_{\text{stat}}} C_{\text{stat}} = (X, a; (Z, G_{\text{bond}})) \\ &\xrightarrow{\tau_{\text{stat}}} I_{\text{stat}} = \big((\mathbf{v}_q, \mathbf{w}_q)_{q \in \mathcal{C}_e}; [v_-, v_+]\big) \xrightarrow{\psi_{\text{stat}}} D \in \mathbb{R}^K, \end{aligned}$$

where $X \in \mathbb{R}^{N \times 3}$ contains the point coordinates and $a \in \mathbb{R}^N$ contains one scalar attribute per point. The optional data $Z = (Z_1, \ldots, Z_N)$ and G_{bond} record the raw atomic numbers of the source points and their chemical bonds, respectively.

The Extraction operator τ_{stat} enumerates geometric events over the point cloud and records them in an ordered set of channels $\mathcal{C}_e$. For a given channel mode, this set and its order are shared by all

molecules in the dataset. For each channel q, $\mathbf{v}_q, \mathbf{w}_q \in \mathbb{R}^{n_q}$ contain the values and corresponding weights of its n_q events. The event counts may vary across channels and molecules. The interval $[v_-, v_+]$ specifies the nominal range of the event values. The Compression operator ψ_{stat} summarizes these channel-wise events into the descriptor D.

H.7.2 STATISTICS-BRANCH FORMATION (κ_{stat})

For each source $s_\ell = (X_\ell, a_\ell)$ in the shared library, we define a corresponding Formation operator κ_{stat} that selects this source and forms the carrier from the source's point coordinates and scalar attributes. It has no additional parameters. For the atomic-number source, the carrier also includes Z and any available G_{bond} to support operators that require chemical information. For all other sources, both are empty.

H.7.3 STATISTICS-BRANCH EXTRACTION (τ_{stat})

The Extraction operator τ_{stat} maps the carrier $C_{\mathrm{stat}} = (X, a; (Z, G_{\mathrm{bond}}))$ to I_{stat}. The selected operator constructs geometric events from the point cloud and assigns their weights and channels according to its parameters.

Let n be the number of events. Their values and weights are represented by $\mathbf{v} = (v_1, \ldots, v_n)$ and $\mathbf{w} = (w_1, \ldots, w_n)$, where (v_e, w_e) is the value and weight of event e.

Table S23 defines the event values v_e for the five operators. For these constructions, let $\mathbf{r}_{ab} = \mathbf{x}_b - \mathbf{x}_a$. The cutoff ϑ defines the neighborhood $\mathcal{N}_c^{\mathrm{stat}}(\vartheta) = \{i \neq c : \|\mathbf{r}_{ci}\|_2 \leq \vartheta\}$, which contains every point other than c when $\vartheta = \infty$. TOR requires the atomic-number source and a usable G_{bond}, and produces no events otherwise.

Table S23: Statistics-branch Extraction operators. Angular event values are in degrees.

Operator	Defining construction
Pair distance (PD)	For every pair $i < j$ with $j \in \mathcal{N}_i^{\mathrm{stat}}(\vartheta)$, compute the distance $v_e = \|\mathbf{r}_{ij}\|_2$.
Triplet angle (TA)	For each center c, consider every pair $i < k$ in $\mathcal{N}_c^{\mathrm{stat}}(\vartheta)$ whose two arms have nonzero length. The event value is the angle $v_e = (180/\pi)\arccos[\mathrm{clip}(\mathbf{r}_{ci}^\top \mathbf{r}_{ck}/(\|\mathbf{r}_{ci}\|_2\|\mathbf{r}_{ck}\|_2), -1, 1)]$.
Radial density (RD)	For each center c and every neighbor $i \in \mathcal{N}_c^{\mathrm{stat}}(\vartheta)$, compute the radial distance $v_e = \|\mathbf{r}_{ci}\|_2$.
Local shape (LS)	For each center c, compute the total second moment, asphericity, and normalized anisotropy: $R_{g,c}^2 = \sum_p \varpi_{cp}$, $\mathrm{asp}_c = \varpi_{c1} - \frac{1}{2}(\varpi_{c2} + \varpi_{c3})$, and $\mathrm{ani}_c = [\sum_p (\varpi_{cp} - R_{g,c}^2/3)^2]^{1/2}/(R_{g,c}^2 + \epsilon)$, where $\varpi_{c1} \geq \varpi_{c2} \geq \varpi_{c3}$ are the eigenvalues of $M_c = \lvert S_c\rvert^{-1} \sum_{i \in S_c} \mathbf{r}_{ci}\mathbf{r}_{ci}^\top$, $S_c = \mathcal{N}_c^{\mathrm{stat}}(\vartheta) \cup \{c\}$.
Strict torsion (TOR)	For every bond (j, k) of G_{bond} with $j < k$ and each choice of bonded neighbors $i \neq k$ of j and $l \notin \{i, j\}$ of k, compute the signed proper dihedral $v_e = (180/\pi)\,\mathrm{atan2}[(\hat{\mathbf{t}}_{jk} \times \mathbf{r}_{ji}^\perp)^\top \mathbf{r}_{kl}^\perp, (\mathbf{r}_{ji}^\perp)^\top \mathbf{r}_{kl}^\perp]$, where $\hat{\mathbf{t}}_{jk} = \mathbf{r}_{jk}/\|\mathbf{r}_{jk}\|_2$ and $\mathbf{v}^\perp = \mathbf{v} - (\mathbf{v}^\top \hat{\mathbf{t}}_{jk})\hat{\mathbf{t}}_{jk}$. Skip the quartet if $\|\mathbf{r}_{jk}\|_2 \leq \epsilon$ or either perpendicular component has norm at most ϵ.

The weighting parameter β is selected from unit, mean, product, and absolute product. For points $i_1, \ldots, i_p$ associated with event e (the center alone for LS), these rules give the event weight w_e as

$$w_e = \begin{cases} 1, & \beta = \mathrm{unit}, \\ p^{-1} \sum_{r=1}^{p} a_{i_r}, & \beta = \mathrm{mean}, \\ \prod_{r=1}^{p} a_{i_r}, & \beta = \mathrm{product}, \\ \lvert \prod_{r=1}^{p} a_{i_r} \rvert, & \beta = \mathrm{absolute\ product}. \end{cases}$$

The channel parameter γ is selected from all, center, neighbor, element-pair, and neighbor-pair. For a pair (v_e, w_e), it determines the assigned channels in $\mathcal{C}_e$. All sources other than the atomic-number source use $\gamma = \mathrm{all}$, which places all events in one channel. The atomic-number source also supports four element-resolved modes, all of which require $\beta = \mathrm{unit}$. For each operator and channel mode, Table S24 identifies the points whose atomic numbers determine the channel labels. Events with the same element or unordered element-pair label are collected in the same channel.

Table S24: Element-resolved channel labels for statistics-branch Extraction operators.

Operator	Center	Neighbor	Element-pair	Neighbor-pair
PD	Z_i	Z_j	(Z_i, Z_j)	(Z_i, Z_j)
TA	Z_c	Z_i, Z_k	(Z_c, Z_i)	(Z_i, Z_k)
RD	Z_c	Z_i	(Z_c, Z_i)	(Z_c, Z_i)
LS	Z_c	Z_c	(Z_c, Z_c)	(Z_c, Z_c)
TOR	Z_j	Z_i, Z_l	(Z_j, Z_k)	(Z_j, Z_k)

Point indices follow Table S23. Pairs in the last two columns are unordered and may contain repeated element labels. In neighbor mode, the two labels of TA or TOR may place the same event in two channels.

For each channel q, the operator collects the assigned event values and weights into $\mathbf{v}_q$ and $\mathbf{w}_q$ in their original event order. These vectors and the nominal range $[v_-, v_+]$ form its output I_{stat}.

H.7.4 Statistics-branch Compression (ψ_{stat})

The Compression operator ψ_{stat} maps I_{stat} to the descriptor $D \in \mathbb{R}^K$. The selected operator m in Table S25 computes a row $Y_{q,:} \in \mathbb{R}^{k_m}$ from the event values and weights $(\mathbf{v}_q, \mathbf{w}_q)$ in each channel of I_{stat}. These rows form $Y \in \mathbb{R}^{|\mathcal{C}_e| \times k_m}$ in channel order, with zero rows for empty channels. The parameters ζ and υ specify the operator's channel pooling and normalization, respectively, giving $D = \mathrm{VNorm}_{\upsilon}^{\mathrm{stat}}(\mathrm{Pool}_{\zeta}(Y))$. In the table, $v_{q,t}$ and $w_{q,t}$ are the value and weight of event t within channel q, and $\odot$ denotes elementwise multiplication. The histogram bin count N_{hist} and the width budget N_{bag} are integers in $\{4, \ldots, 16\}$.

Table S25: Statistics-branch Compression operators.

Operator	Defining construction
Summary statistics (SS)[1]	Summarize the channel events $(\mathbf{v}_q, \mathbf{w}_q)$ by the count n_q, the signed weight sum $\sum_t w_{q,t}$, the weighted mean $\bar{v}_q = \mathcal{Z}_q^{-1} \sum_t w_{q,t} v_{q,t}$, the weighted spread $[\mathcal{Z}_q^{-1} \sum_t \lvert w_{q,t} \rvert (v_{q,t} - \bar{v}_q)^2]^{1/2}$, the unweighted minimum, maximum, and median of $\mathbf{v}_q$, and $\lVert \mathbf{w}_q \odot \mathbf{v}_q \rVert_2$. These eight quantities form $Y_{q,:}$.
Hard histogram (HH)[2]	Partition the range $[v_-, v_+]$ of I_{stat} into N_{hist} equal-width bins $\mathcal{I}_k$. For channel q, accumulate the weight $w_{q,t}$ of each value $v_{q,t}$ in its bin to obtain $Y_{qk} = \sum_t w_{q,t} \mathbb{I}[v_{q,t} \in \mathcal{I}_k]$.
Smoothed histogram (SH)[2]	Represent each event value $v_{q,t}$ by a Gaussian response weighted by $w_{q,t}$. Evaluating their sum at the bin centers c_k of HH gives $Y_{qk} = \sum_t w_{q,t} \exp[-(v_{q,t} - c_k)^2/(2\sigma^2)]$, where the smoothing factor ς sets the SD $\sigma = \varsigma(v_+ - v_-)/N_{\mathrm{hist}}$.
Sorted bag (SB)	Multiply each channel value by its weight to form $\mathbf{w}_q \odot \mathbf{v}_q$. Sort these products in decreasing order and pad or truncate to N_{bag} entries to obtain $Y_{q,:}$.
Channel-preserving bag (CPB)	Apply the SB construction independently to each channel $(\mathbf{v}_q, \mathbf{w}_q)$ to obtain $Y_{q,:} \in \mathbb{R}^{N_{\mathrm{bag}}}$, retaining separate channel representations before pooling by ζ.
Top-k extremes (TK)	Form the weighted channel values $\mathbf{w}_q \odot \mathbf{v}_q$. For a retained count $1 \le \lambda \le N_{\mathrm{bag}}$, concatenate the λ smallest and λ largest entries to obtain $Y_{q,:}$, padding each group independently.
Moment encoding (ME)[1,2]	Rescale each channel value $v_{q,t}$ using the range of I_{stat} as $\widetilde{v}_{q,t} = \mathrm{clip}[(v_{q,t} - v_-)/\max(v_+ - v_-, \epsilon), -10, 10]$. The channel output contains the weighted moments $Y_{qp} = \mathcal{Z}_q^{-1} \sum_t w_{q,t} \widetilde{v}_{q,t}^{\,p}$ for $p = 1, \ldots, \lambda$, where λ is the maximum moment order.
Quantile–CDF (QC)[2]	From the channel values $\mathbf{v}_q$ and absolute weights $\lvert \mathbf{w}_q \rvert$, compute weighted quantiles at $p_k = (k-1)/(N_{\mathrm{bag}} - 1)$ for $k = 1, \ldots, N_{\mathrm{bag}}$ and evaluate the weighted empirical cumulative distribution function (CDF) at $v_- + p_k(v_+ - v_-)$. Concatenate the two vectors to form $Y_{q,:}$. Writing $A_q = \sum_t \lvert w_{q,t} \rvert$, the CDF at x is $\sum_{t: v_{q,t} \le x} \lvert w_{q,t} \rvert / (A_q + \epsilon)$, and the quantiles are unweighted when $A_q \le \epsilon$.

[1] The weight normalizer is $\mathcal{Z}_q = \sum_t \lvert w_{q,t} \rvert + \epsilon$.

[2] The nominal range $[v_-, v_+]$ is $[0, 180]$ for TA, $[-180, 180]$ for TOR, $[0, \max(L_X^2, 1)]$ for LS, and for PD/RD is $[0, \vartheta]$ with finite cutoff ϑ or $[0, \max(L_X, 1)]$ otherwise.

The pooling parameter ζ is selected from row vectorization, sum, mean, SD, mean$\oplus$SD, minimum, maximum, minimum$\oplus$maximum, and median, where $\oplus$ denotes concatenation. Row vectorization concatenates the rows of Y into a vector of length $|\mathcal{C}_e| k_m$. The other choices act componentwise across rows, producing k_m components or $2k_m$ for the two concatenating rules. The normalization parameter υ is selected from identity, ℓ_1, ℓ_2, and max-absolute normalization, in that order. The latter three divide the pooled vector by the corresponding norm, returning zero when that norm vanishes.

H.7.5 ADDRESS DECODING AND ACTIVATION CONDITIONS

For the statistics branch, z_1 selects the Formation operator $\kappa_{\mathrm{stat},\ell}$ with $\ell = 1 + \min\{\lfloor 13 z_1 \rfloor, 12\}$, following the source order in Table S8.

For the Extraction stage, $1 + \min\{\lfloor 5 z_3 \rfloor, 4\}$ selects an operator from Table S23. Only PD, TA, RD, and LS use the neighborhood cutoff, decoded from z_4 as

$$\vartheta = \begin{cases} (0.10 + 0.50 z_4) L_X, & N \geq 15 \text{ and } L_X > \epsilon, \\ \infty, & \text{otherwise.} \end{cases}$$

Every operator decodes the weight parameter β, the entry at position $1 + \min\{\lfloor 4 z_5 \rfloor, 3\}$ of $(\mathrm{unit}, \mathrm{mean}, \mathrm{product}, \mathrm{absolute\ product})$, and the channel parameter γ, the entry at position $1 + \min\{\lfloor 5 z_6 \rfloor, 4\}$ of $(\mathrm{all}, \mathrm{center}, \mathrm{neighbor}, \text{element-pair}, \text{neighbor-pair})$.

For the Compression stage, $1 + \min\{\lfloor 8 z_{11} \rfloor, 7\}$ selects an operator from Table S25. The histogram and bag budgets are $N_{\mathrm{hist}} = \mathrm{round}(4 + 12 z_8)$ and $N_{\mathrm{bag}} = \mathrm{round}(4 + 12 z_9)$. SH sets its smoothing factor to $\varsigma = 0.25 + 2.75 z_{12}$, while TK sets its retained count to $\lambda = \max\{1, \mathrm{round}[1 + z_{12}(N_{\mathrm{bag}} - 1)]\}$. ME sets its maximum moment order to $\lambda = 2 + \mathrm{round}(6 z_{13})$. For every operator, the pooling parameter $\zeta \in \{\text{row vectorization}, \mathrm{sum}, \mathrm{mean}, \mathrm{sd}, \mathrm{mean} \oplus \mathrm{sd}, \mathrm{min}, \mathrm{max}, \mathrm{min} \oplus \mathrm{max}, \mathrm{median}\}$ is selected by the index $1 + \min\{\lfloor 10 z_{10} \rfloor, 9\}$. Indices 1–7 select the first seven options in the listed order. Indices 8 and 9 both select minimum$\oplus$maximum, and index 10 selects median. The normalization parameter $\upsilon \in \{\mathrm{identity}, \ell_1, \ell_2, \text{max-absolute}\}$ is selected by $1 + \min\{\lfloor 4 z_{14} \rfloor, 3\}$ in the listed order.