Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Finite Site Residual Controllability in Molecular QSPR with A Three Gate Test of First Order PDE Identification

Syed Eirfan Atthar

20wj1a03b1@gniindia.org

Dibrugarh University

Research Article

Keywords: Quantitative Structure Property Relationship; Molecular Graph Descriptors; PDE Identifiability; Finite-Site Interpolation; Descriptor Collisions; Alkane Boiling Point

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11661544/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Finite Site Residual Controllability in Molecular QSPR with A Three Gate Test of First Order PDE Identification

Syed Eirfan Atthar* (Corresponding Author)
*Department of Mechanical Engineering, Specialization in Thermal and Fluid, DUIET, Dibrugarh University, Dibrugarh, 786004, India.
Corresponding Author: 20wj1a03b1@gniindia.org
ORCID: 0009-0006-4126-6659

Abstract

This paper develops a three-gate admissibility test for claims that a first-order, constant-coefficient PDE has been inferred from such data. The first gate checks whether the selected descriptors define a single-valued response. The second gives a finite-site construction showing that an interpolant can be changed without altering observed values while prescribing its gradients and, for any nonzero transport vector, its sampled PDE residuals. The third assesses coefficient uniqueness through the rank and singular values of a derivative design matrix, while separately requiring characteristic data to determine the solution. In a 150-alkane benchmark, six graph descriptors produce two collisions with conflicting boiling-point labels. A ridge model gives a mean absolute error of 4.09 ± 0.14 °C under repeated grouped five-fold validation; the pooled leave-one-carbon-number-out mean absolute error is 5.67 °C. Sensitivity across seven Matérn length scales and one cubic radial-basis extension gives candidate coefficient directions separated by as much as 88.03°. The results supports QSPR prediction.

Keywords: Quantitative Structure Property Relationship; Molecular Graph Descriptors; PDE Identifiability; Finite-Site Interpolation; Descriptor Collisions; Alkane Boiling Point.
Mathematics Subject Classification (2020): 35F05; 62J05; 05C09; 05C92.

1 Introduction

A quantitative structure property relationship (QSPR) model relates a numerical representation of molecular structure to an observed physicochemical property. In graph-based QSPR, a molecule is represented by a graph and one or more descriptors are calculated from its vertices, edges, degrees or pairwise distances. Such models range from linear descriptor equations to nonlinear regressors and graph-learning methods. The classical relationship between molecular structure and paraffin boiling points dates at least to Wiener's early use of graph distances, while modern graph-learning methods can learn directly from molecular graphs [1, 2]. More recent QSPR studies use larger descriptor sets, graph-learning methods and ensemble learning to predict molecular properties, including boiling points, across chemically diverse collections [3–6].

A separate research area seeks to discover differential equations from measurements. Sparse identification methods assemble candidate terms from a library and estimate which terms are supported by observations; PDE-discovery procedures typically use spatially and temporally resolved measurements from which derivatives can be estimated [7–9]. Structural identifiability of PDE models is also an active subject, with recent work examining the effects of initial and boundary data on whether model parameters can be distinguished [10, 11]. He, Zhao and Zhong studied how much of a PDE operator can be learned from solution data and showed that recoverability depends on the operator and the available solution and initial data [12]. A recent study also considers the identifiability of state-dependent constitutive functions from sparse, indirect observations of

PDE-governed fields [13]. These studies address recovery from observations of an underlying field or PDE-governed solution. The question here is different and logically prior: whether isolated molecular descriptor–property pairs determine the derivatives needed to assert such an equation in the first place. Their findings reinforce the importance of data structure and initial information, but do not imply that every finite molecular table can be interpreted as a field with observed derivatives.

In a molecular QSPR table, each row normally represents a distinct molecular graph. Its descriptors form a point in a finite subset of a proposed descriptor space, and its measured property is attached to that point. Treating the table as samples from a continuous function is an additional modelling assumption. It requires, at minimum, a well-defined single-valued response, a specified continuous extension, a derivative estimator and sufficient data to distinguish the coefficients of the proposed equation. The distinction is important because the same finite values can be interpolated by different smooth functions whose derivatives do not agree. This is related to the classical theory of differentiable extension from closed sets, but the present question concerns the consequences for first-order PDE claims made from molecular descriptor tables [14].

The issue is also relevant to topological-index QSPR studies. Recent studies have already used Zagreb-vector and VL indices in regression models for melting and boiling points of aliphatic hydrocarbons, and Zagreb-vector and total-eccentricity descriptors for physical properties of linear alkanes [15, 16]. Descriptor degeneracy has also been identified directly as a difficulty in QSPR: Song, Wen and Chai developed a graph-signal-processing framework with a higher-level compound graph to create more discriminating features for compounds that conventional molecular descriptors cannot distinguish [17]. That work is particularly relevant to the collision check used here. The present study does not claim descriptor degeneracy as a new phenomenon or propose a new descriptor to resolve it. Instead, it examines the additional consequence for differential-equation inference: even after a collision-free set of sites has been formed or an explicit aggregation rule has been applied, the finite property values do not determine the first derivatives required by a PDE. Recent QSPR work using learned molecular descriptors and graph-theoretic indices reinforces the need to distinguish predictive representation from a differential law [5, 6]. Therefore, neither the general use of topological descriptors for alkane properties nor the use of regression for boiling-point prediction is claimed as new here.

The contribution of this work is a chemistry-specific test for deciding whether a finite descriptor table supports a claim of first-order, constant-coefficient PDE identification. It combines three logically separate checks. First, exact descriptor collisions are examined because identical inputs with different measured outputs cannot be represented by one deterministic function. Second, a finite-site construction establishes that property values do not determine the descriptor gradients or the sampled residuals of a chosen first-order equation. Third, after a continuous extension has been fixed, a rank and singular-value analysis tests whether the PDE coefficients are conditionally identifiable; sufficient characteristic data are then required to determine the solution. A numerical extension-sensitivity experiment and an expanded alkane QSPR benchmark illustrate the practical consequences.

The literature positioning is intentionally bounded. A targeted search conducted on 7 November 2026 used combinations of terms relating to data-driven PDE discovery, structural identifiability, molecular descriptors, descriptor degeneracy, graph-signal processing, QSPR and alkane boiling-point prediction. It located substantial prior work in each of those areas, including descriptor-degeneracy remedies in QSPR, but did not locate a study applying the same sequence of descriptor-consistency, finite-site residual-controllability and extension-conditioned rank checks to molecular QSPR data. This search was not a systematic review and cannot establish that no such study exists. The paper does not propose a new general theorem of smooth interpolation, a new molecular descriptor, or a discovered governing law for alkane boiling points. Its contribution is the formulation and numerical examination of a chemistry-oriented admissibility protocol that separates prediction from differential-law identification. The interpolation construction is

elementary and belongs to the broader tradition of finite-set extension theory; its role here is to make the implications for molecular PDE claims explicit and testable.

Let G_i denote the molecular graph for compound i, let $z_i \in \mathbb{R}^p$ be its descriptor vector and let $y_i \in \mathbb{R}$ be the measured property. The observed dataset is

$$\mathcal{D} = \{(z_i, y_i) : i = 1, \dots, N\}. \tag{1.1}$$

A deterministic continuous QSPR representation is a function $u : \Omega \subseteq \mathbb{R}^p \to \mathbb{R}$ satisfying $u(z_i) = y_i$ at the observed sites, either exactly or within a specified measurement model.

Proposition 1.1 (Descriptor-collision obstruction)**.** *If two observations satisfy $z_i = z_j$ but $y_i \neq y_j$, no single-valued function $u : \Omega \to \mathbb{R}$ can interpolate both observations exactly.*

Proof. A function assigns one value to each argument. The equality $z_i = z_j$ therefore implies $u(z_i) = u(z_j)$, which is incompatible with $y_i \neq y_j$. □

This obstruction does not prevent statistical prediction. A regression model may estimate a conditional mean or another summary of a response distribution. It does mean that an exact deterministic surface over the selected descriptors is not defined by the conflicting rows without an explicit rule for replicates, noise or aggregation.

For squared-error prediction, the same collision check gives a simple lower bound. Let $\mathcal{G}$ be the set of groups that share an identical descriptor vector, and let $\bar{y}_g$ denote the mean response in group g. Any deterministic predictor that depends only on the selected descriptor vector must assign one value to every member of a group, so its total squared error over that group is at least $\sum_{i \in g}(y_i - \bar{y}_g)^2$. Consequently, for the full dataset,

$$\mathrm{RMSE} \;\geq\; \left[\frac{1}{N}\sum_{g \in \mathcal{G}}\sum_{i \in g}(y_i - \bar{y}_g)^2\right]^{1/2}. \tag{1.2}$$

This is a lower bound caused only by descriptor collisions; it does not include measurement noise or model approximation error.

Suppose first that the descriptor sites $z_1, \dots, z_N$ are pairwise distinct. For a proposed first-order, homogeneous, constant-coefficient equation, write

$$a^{\mathsf{T}}\nabla u(z) + c\,u(z) = 0, \qquad a \in \mathbb{R}^p, \quad c \in \mathbb{R}. \tag{1.3}$$

The table in equation (1.1) contains values of u at the observed sites, but does not directly contain the gradient vectors $\nabla u(z_i)$. The following result states more than the fact that those gradients are unknown: they can be prescribed without changing any of the observed function values.

Theorem 1.2 (Finite-site residual controllability)**.** *Let $z_1, \dots, z_N$ be pairwise distinct points in $\mathbb{R}^p$, let $y_1, \dots, y_N \in \mathbb{R}$, and let $a \in \mathbb{R}^p \setminus \{0\}$ and $c \in \mathbb{R}$ be fixed. For any prescribed residual vector $r = (r_1, \dots, r_N)^{\mathsf{T}} \in \mathbb{R}^N$, there exists a polynomial $u : \mathbb{R}^p \to \mathbb{R}$ of total degree at most $2N - 1$ such that*

$$u(z_i) = y_i, \qquad a^{\mathsf{T}}\nabla u(z_i) + c\,u(z_i) = r_i, \quad i = 1, \dots, N. \tag{1.4}$$

Proof. For each site define the polynomial

$$P_i(z) = \prod_{j \neq i} \frac{\|z - z_j\|^2}{\|z_i - z_j\|^2}. \tag{1.5}$$

Then $P_i(z_i) = 1$ and $P_i(z_j) = 0$ for $j \neq i$. Hence

$$u_0(z) = \sum_{i=1}^{N} y_i P_i(z) \tag{1.6}$$

is a polynomial interpolant of the observed values. Define the residual of u_0 at site z_i by

$$s_i = a^{\mathsf{T}} \nabla u_0(z_i) + c y_i. \tag{1.7}$$

Since $a \neq 0$, the vectors

$$q_i = \frac{r_i - s_i}{\|a\|^2} a \tag{1.8}$$

are well-defined. Now set

$$h_i(z) = \prod_{j \neq i} \|z - z_j\|^2, \qquad g(z) = \sum_{i=1}^{N} \frac{h_i(z)}{h_i(z_i)} q_i^{\mathsf{T}}(z - z_i). \tag{1.9}$$

The sites are distinct, so $h_i(z_i) > 0$. At z_i, the ith summand in g has value zero and gradient q_i. At every other site z_j, the factor $\|z - z_j\|^2$ makes both the summand and its first derivative vanish. Consequently, $g(z_i) = 0$ and $\nabla g(z_i) = q_i$ for all i. The polynomial $u = u_0 + g$ retains the observations and satisfies

$$\begin{aligned} a^{\mathsf{T}} \nabla u(z_i) + c u(z_i) &= s_i + a^{\mathsf{T}} q_i \\ &= s_i + (r_i - s_i) = r_i. \end{aligned}$$

The polynomial u_0 has degree at most $2(N-1)$, while each summand in g has degree at most $2(N-1)+1$. Hence $u = u_0 + g$ has total degree at most $2N-1$. This proves the result. □

Corollary 1.3 (A finite table cannot validate a sampled PDE by residual fit alone)**.** *Under the hypotheses of the theorem, for any selected $a \neq 0$ and c, there is a smooth interpolant that satisfies equation (1.3) exactly at every observed descriptor site. There is also an interpolant with any other prescribed residual vector at those sites.*

Proof. Choose $r_i = 0$ for all i to obtain exact satisfaction at the observed sites. The general case follows directly from the theorem. □

The theorem concerns values and first derivatives at a finite collection of sites. It does not claim that every such interpolant satisfies the PDE throughout a continuous region. Rather, it shows why an equation that has been fitted only at the observed molecular points cannot be regarded as identified from the table: the same values permit arbitrary sampled residuals unless the admissible extension or derivative field is constrained independently. The construction is consistent with the broader theory of differentiable extension from closed sets [14].

Suppose that a continuous extension u and its derivatives have been specified. At evaluation sites $\widetilde{z}_1, \ldots, \widetilde{z}_M$, define the derivative design matrix

$$A_u = \begin{bmatrix} \partial_{z_1} u(\widetilde{z}_1) & \cdots & \partial_{z_p} u(\widetilde{z}_1) & u(\widetilde{z}_1) \\ \vdots & & \vdots & \vdots \\ \partial_{z_1} u(\widetilde{z}_M) & \cdots & \partial_{z_p} u(\widetilde{z}_M) & u(\widetilde{z}_M) \end{bmatrix} \in \mathbb{R}^{M \times (p+1)}. \tag{1.10}$$

The sampled equation is $A_u \theta = 0$, where $\theta = (a_1, \ldots, a_p, c)^{\mathsf{T}}$. A nonzero coefficient vector is defined only up to multiplication by a nonzero scalar.

Proposition 1.4 (Rank criterion)**.** *For a fixed derivative field, the homogeneous system $A_u \theta = 0$ has a one-dimensional nonzero null space, and hence determines one coefficient direction up to scale, if and only if*

$$\operatorname{rank}(A_u) = p. \tag{1.11}$$

If $\operatorname{rank}(A_u) = p+1$*, only the zero coefficient vector satisfies the sampled equations. If* $\operatorname{rank}(A_u) \leq p-1$*, more than one independent coefficient direction satisfies them. The rank condition alone does not guarantee a transport term: the unique null vector must also have a nonzero first-p-component block, that is, $\|a\| > 0$.*

Proof. The matrix A_u has $p+1$ columns. By the rank–nullity theorem,

$$\dim \ker(A_u) = p + 1 - \operatorname{rank}(A_u). \tag{1.12}$$

The null space is one-dimensional precisely when its rank is p. A full column rank matrix has a trivial null space, and a rank of at most $p-1$ gives a null space of dimension at least two. □

For noisy derivatives, an exact null vector will generally not exist. A common numerical candidate is the right singular vector associated with the smallest singular value of A_u. Such a vector is interpretable only when its direction is stable under changes in the extension, derivative estimator, sampling sites and data perturbations. The singular-value gap and the residual norm should therefore be reported alongside a coefficient vector. The rank criterion is conditional on the selected extension; it does not resolve the non-identifiability established by the previous theorem.

Even when the coefficient direction in equation (1.3) is known, the equation does not determine a unique solution without suitable initial or boundary information. For example, consider

$$u_x + \alpha u_y + \beta u_z - \gamma u = 0. \tag{1.13}$$

Its characteristic representation is

$$u(x, y, z) = e^{\gamma x} F(y - \alpha x, z - \beta x), \tag{1.14}$$

where F is an arbitrary function of two variables. A single value of u fixes only one value of F. A suitable initial condition on a non-characteristic two-dimensional surface is needed to determine the solution over the characteristic domain. This requirement is separate from coefficient identification.

The results yield a practical test for a proposed PDE-based QSPR study:

1. **Descriptor consistency.** Check exact repeated descriptor vectors. Conflicting outputs require a declared noise, aggregation or probabilistic model before a single-valued response surface can be assumed.
2. **Extension and derivative specification.** State the continuous domain, the interpolation or smoothing family, its hyperparameters and the derivative estimator. Test whether gradients and PDE residuals change materially under reasonable alternatives.
3. **Coefficient and solution checks.** Report matrix rank, singular values, the null-space dimension and coefficient stability. Supply suitable characteristic data and test predictions on held-out molecules or chemical series.

Failure at an early gate does not invalidate QSPR regression. It means that a differential-law claim is not supported by the available information.

2 Data and computational methods

The five-by-five synthetic grid in the initial manuscript is generated by

$$u(x, t) = \frac{2x}{1 + 2t}, \qquad 0 \le x \le 1, \quad 0 \le t \le 1. \tag{2.1}$$

Its derivatives are

$$u_x = \frac{2}{1 + 2t}, \qquad u_t = -\frac{4x}{(1 + 2t)^2}. \tag{2.2}$$

Direct substitution gives

$$u_t + u u_x = 0. \tag{2.3}$$

This identity is a particular solution of the inviscid Burgers equation. It is used here only to audit the synthetic table, not as a new physical model. The original manuscript instead assigned the

grid to $u_t - 8u_x + 50u = 0$. Both residual fields were evaluated using the exact derivatives at the 25 displayed grid points.

The preliminary table contains 11 normal alkanes from methane to *n*-undecane. Its two reported topological columns are labelled MV1 and VL, but the source manuscript does not define either index mathematically or provide a calculation procedure. Their values therefore cannot be independently reproduced from the manuscript. All 11 reported normal boiling points were compared with the recommended values in the February 2026 NIST compilation; fusion temperatures were checked selectively against the NIST Chemistry WebBook. Temperatures were converted using $T_{^\circ\mathrm{C}} = T_{\mathrm{K}} - 273.15$.

Table 1: Selected checks of physical-property values in the original 11-alkane table. NIST entries are identified as recommended values or individual WebBook measurements

Compound	Property	Original (°C)	NIST comparison and assessment
Methane	Normal boiling	−164.00	NIST 2026 recommended value 111.667 ± 0.013 K (−161.483 ± 0.013 °C); the working value is lower by about 2.52 °C.
Propane	Fusion	−183.00	Selected value 85.5 ± 0.2 K (−187.65 ± 0.20 °C); verify phase definition and source before replacement.
n-Butane	Fusion	−38.00	134.72–134.81 K (−138.43 to −138.34 °C); discrepancy is approximately 100.4 °C.
n-Undecane	Fusion	−10.36	NIST value 247.54 K (−25.61 °C); discrepancy is approximately 15.25 °C.
n-Undecane	Normal boiling	156.58	NIST 2026 recommended value 469.05 ± 0.15 K (195.900 ± 0.150 °C); the working value is lower by about 39.32 °C.

A separate comparison of all 11 boiling-point entries against the recommended values in the February 2026 NIST compilation shows that the methane entry is lower by about 2.52 °C, the ethane-to-decane entries differ by no more than about 0.51 °C, and the *n*-undecane entry is lower by about 39.32 °C [18]. The propane fusion value differs from one NIST-listed measurement and should be checked against the original source and phase convention before being replaced [19]. The *n*-butane and *n*-undecane fusion entries also show major discrepancies [20, 21]. NIST cautions that pressure-derived uncertainty for the normal boiling points of *n*-undecane, *n*-dodecane and *n*-hexadecane may be around 0.5–1.5 K rather than about 0.1 K. That range remains small beside the 39.32 °C discrepancy in the working table [18].

The boiling-point comparison covers all 11 entries. The fusion-point check is selective because the original manuscript gives no measurement sources or phase conventions. The preliminary table therefore cannot support a chemical law without a complete provenance record.

The two reported descriptor columns in the original table are also almost perfectly correlated: the Pearson correlation computed from the displayed values is $r = 0.999737$, and the condition number of the standardized two-column matrix is 87.25. This collinearity makes it difficult to separate their individual statistical contributions in an 11-point model. It does not provide an estimate of a partial derivative.

The expanded QSPR benchmark uses the University of Pisa collection of 150 acyclic alkanes, covering carbon numbers one through ten and 1331 carbon vertices in total [2, 22]. The dataset description explicitly states that target values are stored as boiling-point temperatures in degrees Celsius divided by 100; the values were multiplied by 100 to express them in degrees Celsius [23]. This interpretation is consistent with the normal boiling points of the smallest alkanes and was cross-checked against NIST records.

The molecular encodings were parsed as rooted trees. Eight source records contained one surplus closing parenthesis at the end of the encoding. In those cases, only that terminal character was removed after a strict parse failed; no internal token was changed. The parser checked that all

150 identifiers were present, that the reconstructed graphs contained 1331 vertices, that each graph had $|E| = |V| - 1$, and that the handshaking identity $\sum_{v \in V} d(v) = 2|E|$ held. The maximum degree did not exceed the carbon valence bound of four. The class counts for $n_C = 1, \ldots, 10$ were 1, 1, 1, 2, 3, 5, 9, 18, 35 and 75, respectively. The computational analysis used Python 3.13.5, NumPy 2.3.5, pandas 2.2.3, scikit-learn 1.8.0 and Matplotlib 3.10.8; the exact environment and model settings are recorded in the supplementary validation file.

For each carbon-skeleton tree $G = (V, E)$, six descriptors were calculated:

$$n_C(G) = |V|, \tag{2.4}$$

$$W(G) = \sum_{\{v,w\} \subseteq V} d_G(v, w), \tag{2.5}$$

$$M_1(G) = \sum_{v \in V} d(v)^2, \tag{2.6}$$

$$M_2(G) = \sum_{\{v,w\} \in E} d(v)d(w), \tag{2.7}$$

$$B(G) = |\{v \in V : d(v) \geq 3\}|, \tag{2.8}$$

$$\Delta(G) = \max_{v \in V} d(v), \tag{2.9}$$

where $d(v)$ is the vertex degree and $d_G(v, w)$ is the shortest-path distance. The Wiener index is the sum of distances over unordered vertex pairs; M_1 and M_2 are the first and second Zagreb indices. These are established graph descriptors; no novelty is claimed for their definitions [1].

A descriptor collision occurs when different molecular graphs share exactly the same six-dimensional descriptor vector. The 150 records contain 148 distinct descriptor vectors and two repeated groups. The repeated groups are shown in table 2. Both groups have different boiling-point labels. Therefore, the six-descriptor map is not an exact single-valued representation of all 150 source responses. Applying equation (1.2) gives a collision-only RMSE lower bound of 0.0603 °C for a deterministic predictor evaluated on all 150 records. This floor is small relative to the observed cross-validation errors; its significance is the failure of exact single-valued interpolation, not a large contribution to predictive error. For the QSPR validation, identical descriptor vectors were assigned to the same outer fold so that a duplicate input could not occur in both training and test partitions. For the extension-sensitivity experiment only, the two conflicting responses at each repeated site were averaged to obtain one response value per unique descriptor vector.

Table 2: Exact collisions in the six-descriptor representation of the 150 alkane records. Descriptor order is $(n_C, W, M_1, M_2, B, \Delta)$

Record IDs	Number of records	Descriptor vector	Source boiling points (°C)
86, 103	2	$(10, 122, 42, 44, 2, 4)$	159.0, 158.0
136, 139	2	$(10, 138, 38, 38, 2, 3)$	160.0, 159.7

Three regression models were compared. The first was ordinary least squares (OLS) using carbon number alone. The second was ridge regression using the six descriptors after standardization. Ridge regression was selected because several descriptors are correlated; the penalty reduces coefficient instability in non-orthogonal predictor sets [24]. Its penalty parameter was selected from 25 values on a logarithmic grid, 10^{-3} to 10^4. The third model was an extremely randomized tree regressor with 300 trees, a minimum leaf size of two, all features considered at each split and random seed 42. The tree settings were fixed in advance and were not tuned on the held-out data.

The first evaluation used five repetitions of shuffled five-fold cross-validation, with outer random seeds 2026–2030. Outer folds were grouped by exact six-descriptor vector, so repeated descriptor rows were kept together. In every outer training set, ridge tuning used three-fold grouped inner

validation. Scaling and model fitting were repeated separately within each inner training fold; the outer test fold was not used for scaling or penalty selection. One out-of-fold prediction was generated for every record in each repetition. The reported mean and standard deviation are calculated across the five complete out-of-fold repetitions.

The second evaluation was leave-one-carbon-number-out (LOCO) validation. In each of ten outer folds, all molecules with one fixed carbon number were held out together. Ridge tuning within the remaining compounds used five-fold grouped inner validation by carbon number, with all scaling estimated inside each inner training fold. The pooled LOCO metrics were calculated over all 150 predictions. The macro-group MAE is the arithmetic mean of the ten per-carbon-number MAEs and gives each carbon-number class equal weight. The class sizes are highly imbalanced, and the C1–C3 groups contain one compound each; their individual error estimates are consequently unstable.

The descriptor collision check shows that no exact single-valued interpolant passes through all 150 original pairs. To assess the derivative sensitivity of a specified family of extensions, the response values within each collision group were averaged, leaving 148 unique descriptor sites. The six descriptors were standardized using the means and standard deviations of these 148 sites. The kernel system was solved using the centred response values and the response mean was restored in the interpolant. For the PDE design matrix, the mean-restored property field and its gradient were both divided by the response standard deviation, without subtracting the response mean. This scaling preserves the homogeneous equation form; subtracting an additive constant would change the reaction term.

A Matérn 3/2 radial basis kernel, a standard smoothness-controlled choice in scattered-data interpolation, was used [25]:

$$k_\ell(r) = \left(1 + \frac{\sqrt{3}r}{\ell}\right) \exp\left(-\frac{\sqrt{3}r}{\ell}\right), \tag{2.10}$$

where r is Euclidean distance in the standardized descriptor coordinates and ℓ is a length scale. For the 148 sites, let K_ℓ be the kernel matrix and let $\eta = 10^{-9}$ be a small nugget. The interpolant was

$$u_\ell(z) = \bar{y} + k_\ell(z, Z)^\mathsf{T}(K_\ell + \eta I)^{-1}(y - \bar{y}\mathbf{1}), \tag{2.11}$$

where $\bar{y}$ is the mean response at the unique sites. For a site z_j, the analytic derivative used in the gradient matrix is

$$\nabla_z k_\ell(\|z - z_j\|) = -\frac{3}{\ell^2} \exp\left(-\frac{\sqrt{3}\,\|z - z_j\|}{\ell}\right)(z - z_j). \tag{2.12}$$

The derivatives were used to form A_{u_ℓ/s_y} in equation (1.10), where s_y is the standard deviation of the 148 site responses. Both the derivative columns and the final response column were divided by s_y; the mean-restored response was not mean-centred in this matrix because an additive shift would change the homogeneous reaction term. The gradient implementation was independently checked by central finite differences.

To test whether the observed sensitivity was peculiar to the Matérn family, a second extension was fitted using the cubic polyharmonic radial basis function $\phi(r) = r^3$ with an affine polynomial tail. Its coefficients solve the augmented interpolation system

$$\begin{bmatrix} K & P \\ P^\mathsf{T} & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ b \end{bmatrix} = \begin{bmatrix} (y - \bar{y}\mathbf{1})/s_y \\ 0 \end{bmatrix}, \qquad P = [\mathbf{1}, Z], \tag{2.13}$$

where $K_{ij} = \|z_i - z_j\|^3$ and $b = (b_0, b_1, \ldots, b_6)^\mathsf{T}$. The gradient of the fitted field, divided by s_y, is

$$\nabla u_{\mathrm{cub}}(z_i)/s_y = \sum_{j=1}^{148} 3\lambda_j \|z_i - z_j\|\,(z_i - z_j) + (b_1, \ldots, b_6)^\mathsf{T}. \tag{2.14}$$

The affine tail is required for the conditionally positive-definite cubic radial basis interpolant [26]. The augmented interpolation system and its conditioning are reported separately from the Matérn kernel matrices. This additional fit is used as a second, mathematically distinct extension family rather than as a preferred chemical interpolation.

The length scales were $\ell \in \{0.35, 0.5, 0.75, 1.0, 1.5, 2.0, 3.0\}$. At each scale, a singular value decomposition of $A_{u\ell}$ was used to obtain the smallest right singular vector as a least-squares candidate coefficient direction. The sign was chosen so that the final component was non-negative. Candidate directions were compared using the sign-invariant angle

$$\theta_{ij} = \cos^{-1}\left(\left|v_i^{\mathsf{T}} v_j\right|\right), \tag{2.15}$$

where all vectors were computed in the same standardized coordinates. This angle spread is a descriptive sensitivity summary for the selected kernel family, not a universal identifiability metric. The numerical rank was computed using the standard singular-value tolerance implemented by NumPy.

3 Results

The displayed synthetic data agree with equation (2.1). Across the 25 grid points, the maximum absolute residual of $u_t + uu_x$ is 1.11×10^{-16}, which is at floating-point round-off. The maximum absolute residual of the linear equation stated in the original formulation, $u_t - 8u_x + 50u$, is 80.0 and its root-mean-square residual is 29.59. The corrected nonlinear identity is therefore consistent with the table; the original linear equation is not. The two residual fields are shown in figure 1.

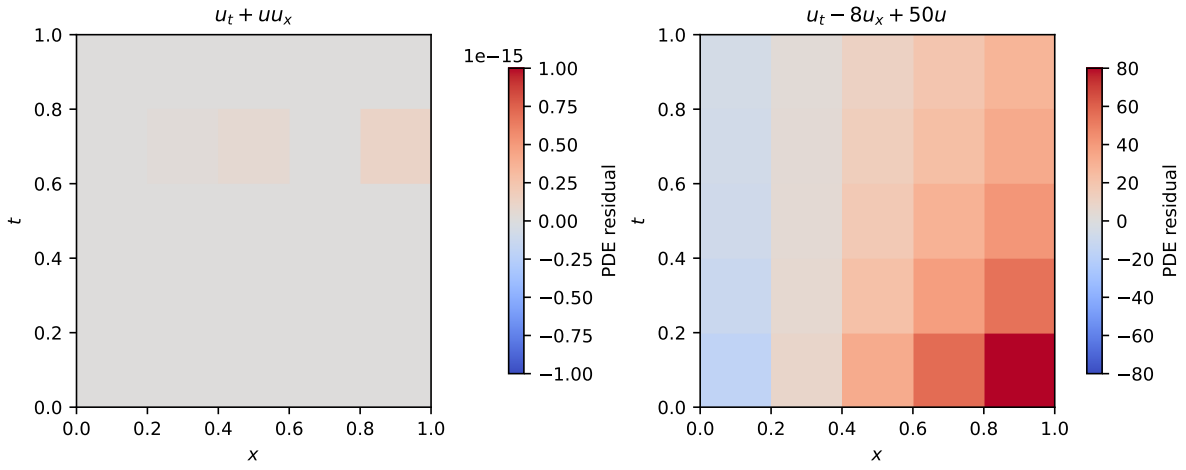


Fig. 1 Residual fields on the original five-by-five (x, t) grid. The left panel shows the residual of $u_t + uu_x = 0$ for the field in equation (2.1); the right panel shows the residual of $u_t - 8u_x + 50u = 0$. The left panel is at floating-point round-off, whereas the right panel is materially non-zero

The selected physical-property checks in table 1 identify major inconsistencies in the original 11-alkane working table, particularly the *n*-butane melting point and the *n*-undecane melting and boiling points. The methane boiling-point entry is lower than the February 2026 NIST recommended value by about 2.52 °C; the difference is larger than the stated NIST uncertainty, although the provenance of the original working value is not given. The propane fusion entry requires a phase-specific source check. Since the MV1 and VL indices are not defined in the original manuscript, their listed values cannot be reproduced or validated from that document.

The public benchmark contains 150 records, but only 148 unique six-descriptor vectors. The two conflicting groups in table 2 establish that the six descriptors do not provide an exact one-to-one response representation for every listed molecule. The descriptor correlation matrix is

shown in figure 3. The correlation between M_1 and M_2 is 0.984; carbon number is also strongly correlated with the Wiener index (0.927), M_1 (0.942) and M_2 (0.907). The condition number of the standardized six-feature matrix is 25.04. These associations do not invalidate ridge regression, but they limit the interpretation of individual regression coefficients as separate physical effects.

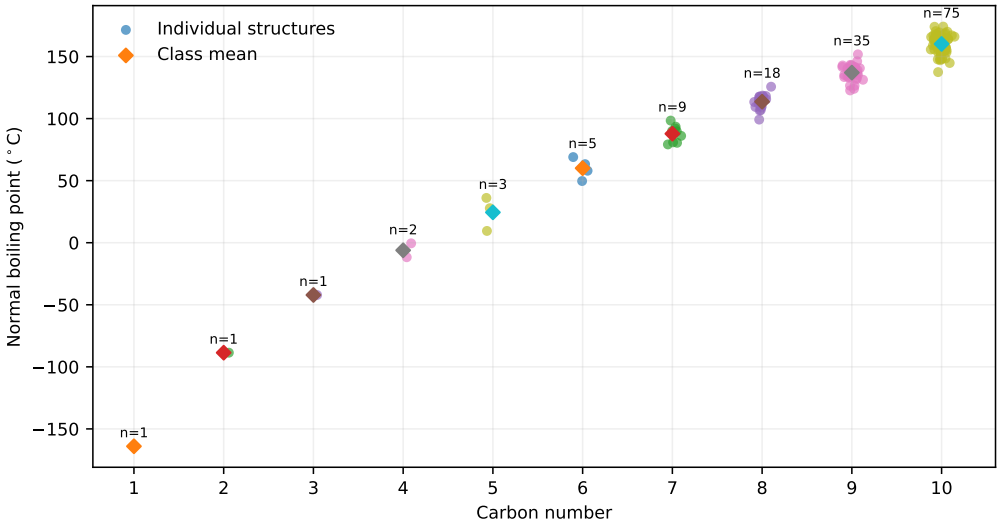


Fig. 2 Normal boiling-point responses in the 150-alkane benchmark. Points are individual structures; diamonds show the mean for each carbon-number class. The class sizes are annotated above the groups

The six-descriptor ridge model produced a mean MAE of 4.09 ± 0.14 °C in repeated grouped five-fold validation, compared with 7.23 ± 0.20 °C for carbon-number OLS and 5.53 ± 0.36 °C for ExtraTrees. The ridge model also had the lowest mean RMSE and the highest mean R^2 in this evaluation. The summary is given in table 3; observed and averaged out-of-fold predictions are shown in figure 4.

Table 3: Repeated grouped five-fold validation. Values are the mean ± standard deviation across five complete repetitions. All preprocessing and ridge penalty selection were performed within the relevant training folds

Model	MAE (°C)	RMSE (°C)	R^2
Carbon-number OLS	7.23 ± 0.20	11.11 ± 0.42	0.9479 ± 0.0040
Six-descriptor ridge	4.09 ± 0.14	6.84 ± 0.84	0.9801 ± 0.0049
ExtraTrees	5.53 ± 0.36	12.59 ± 1.90	0.9319 ± 0.0201

The improvement over carbon number alone indicates that the selected graph descriptors contain predictive information within this dataset. It does not establish that each descriptor has an independently identifiable effect or that a differential law governs the boiling point. The descriptors are correlated, and the two descriptor collisions show that their combined representation is still incomplete for exact interpolation.

When an entire carbon-number class was held out, the six-descriptor ridge model gave a pooled MAE of 5.67 °C and a pooled RMSE of 8.05 °C. The macro-average of the ten group-specific MAEs was 9.61 °C. Carbon-number OLS gave a pooled MAE of 10.75 °C and a macro-group MAE of 18.53 °C, while ExtraTrees gave 15.57 and 19.53 °C, respectively. Results are summarised in table 4.

The pooled MAE is influenced by the larger classes, especially C8–C10, whereas the macro-group value gives the singleton C1–C3 classes the same weight as the much larger C10 class. The ridge model's C1 group error is 52.57 °C, reflecting the difficulty of predicting the isolated methane

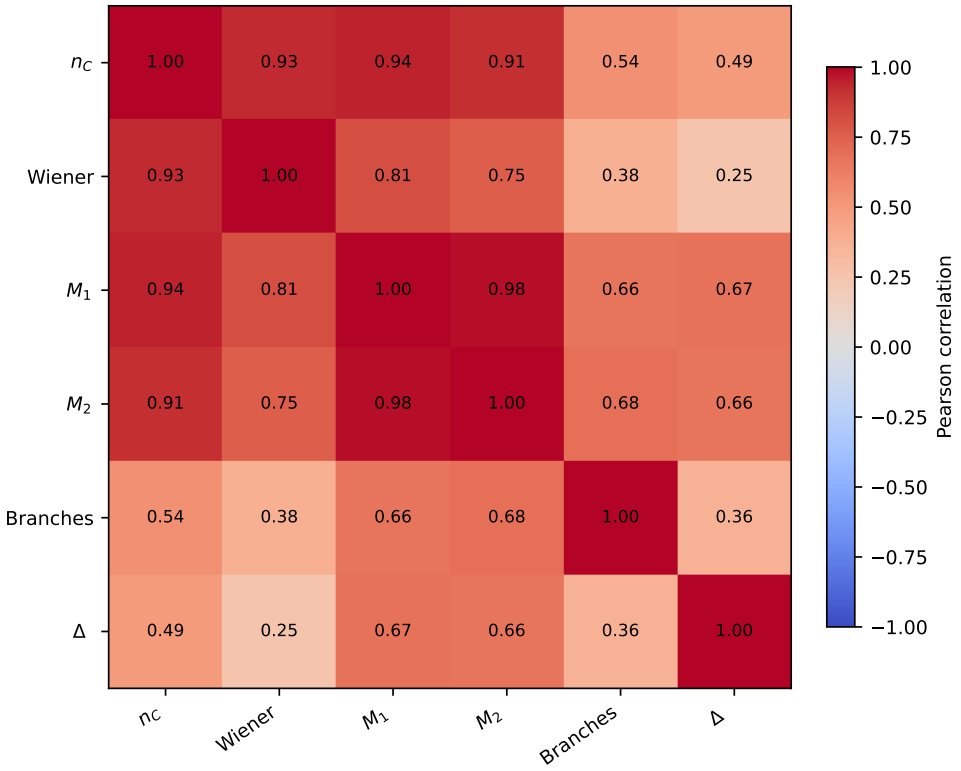


Fig. 3 Pearson correlation matrix for the six graph descriptors used in the QSPR benchmark. The strongest pairwise association is between the first and second Zagreb indices

Table 4: Leave-one-carbon-number-out validation. Each outer test fold contains every molecule from one carbon-number class. Pooled metrics weight compounds equally; the macro-group MAE weights the ten classes equally

Model	Pooled MAE (°C)	Pooled RMSE (°C)	Pooled R^2	Macro-group MAE (°C)
Carbon-number OLS	10.75	13.97	0.9177	18.53
Six-descriptor ridge	5.67	8.05	0.9727	9.61
ExtraTrees	15.57	20.21	0.8278	19.53

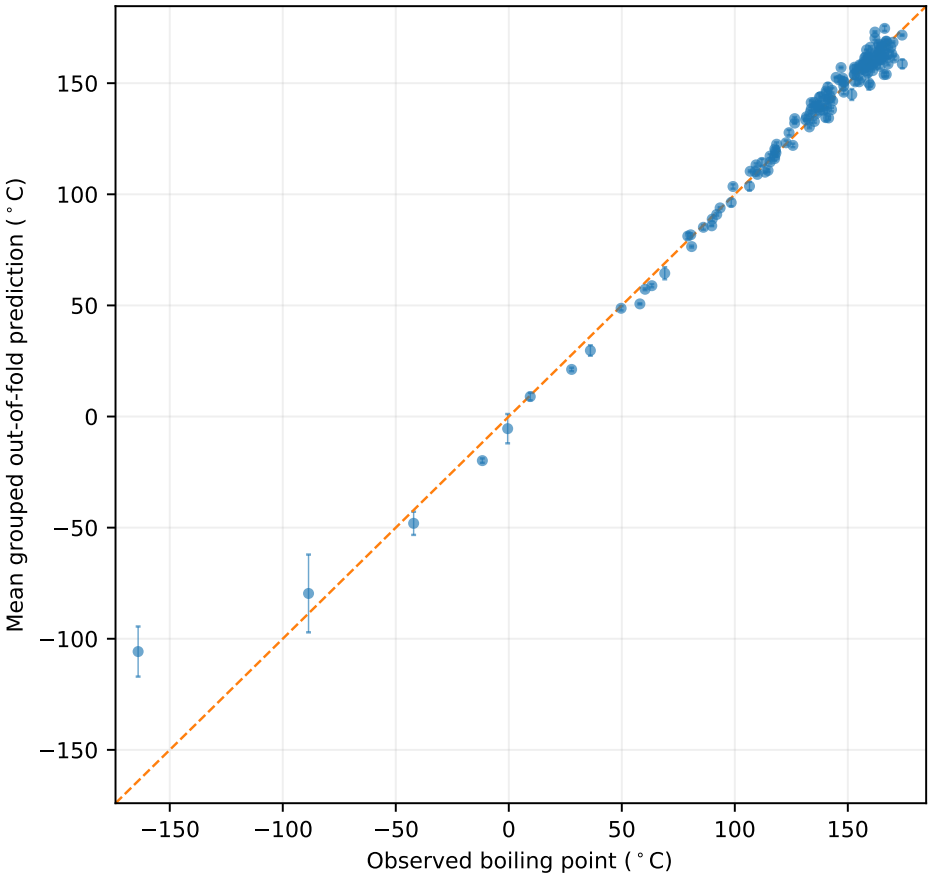


Fig. 4 Observed and mean out-of-fold predicted boiling points for the six-descriptor ridge model. Predictions are averaged over the five grouped cross-validation repetitions; error bars indicate the standard deviation of the repeated predictions for each compound

point when the entire C1 class is withheld. The high pooled R^2 should therefore not be read in isolation. The validation comparison is displayed in figure 5.

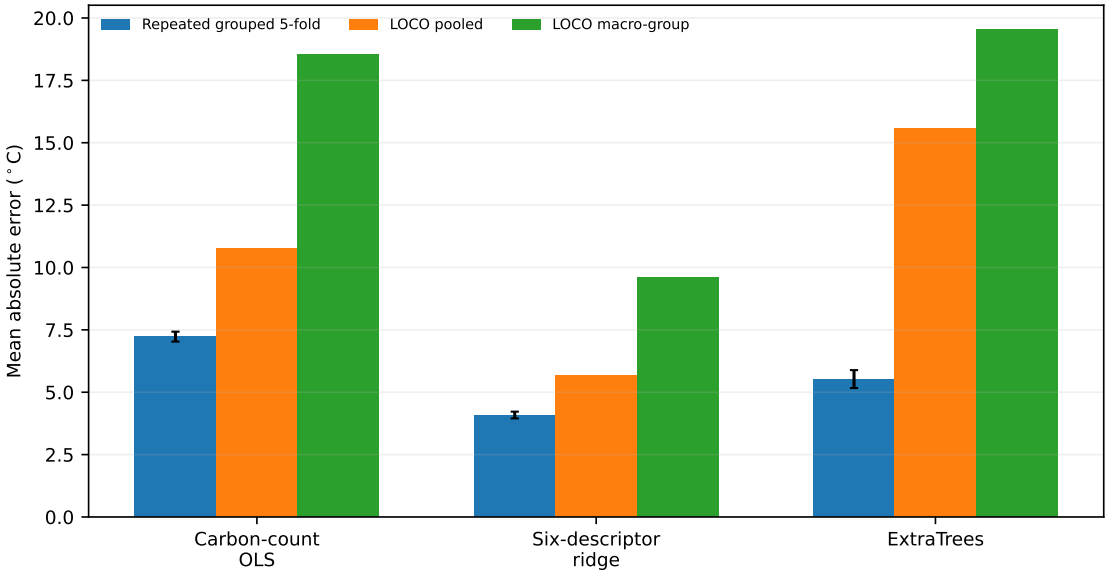


Fig. 5 MAE for the three QSPR models under repeated grouped five-fold validation and leave-one-carbon-number-out validation. LOCO is shown as a pooled score and as the unweighted mean of the ten group-specific MAEs

For each of the seven Matérn length scales, the RBF interpolant reproduced the averaged response at the 148 unique descriptor sites with an interpolation RMSE between 8.45×10^{-7} and 5.18×10^{-4} °C. Nevertheless, the computed derivative design matrix had numerical rank seven at every scale under the stated numerical tolerance. Since it has seven columns ($p + 1$ for $p = 6$), none of these computed matrices had a nontrivial numerical null space at that tolerance. Thus, the selected interpolants did not yield an exact coefficient vector at the working numerical precision; only a least-squares candidate direction could be reported.

For the seven Matérn fits, the ratio σ_6/σ_7 ranged from 1.087 to 3.194, indicating that the smallest singular direction was not strongly separated from its neighbour at most length scales. The cubic polyharmonic interpolant also produced a numerically full-rank derivative design matrix at the same tolerance. Its smallest right singular vector is therefore an approximate least-squares direction, not an exact coefficient vector. Its interpolation RMSE was 1.82×10^{-8} °C, with $\sigma_6/\sigma_7 = 1.385$. Its augmented interpolation system had a condition number of 1.63×10^9. Across all eight candidate directions (seven Matérn length scales and the cubic interpolant), the pairwise sign-invariant angles ranged from 6.13° to 88.03°, with a median of 61.27°. The largest angle shows that very close fits to the same observed values can yield substantially different least-squares PDE coefficient directions even when a second radial-basis family is considered. The Matérn kernel-matrix condition numbers ranged from 2.29×10^4 to 1.13×10^8. These results are shown in figure 6 and table 5.

Table 5: Sensitivity of the sampled PDE design matrix to the continuous-extension choice. The RMS residual is dimensionless because both response values and gradients are divided by the response standard deviation. For Matérn rows, the condition number is that of $K_\ell + \eta I$; for the cubic row, it is that of the augmented interpolation system in equation (2.13). These system condition numbers are not directly comparable.

Family	Parameter	Interpolation RMSE (°C)	System κ	σ_6/σ_7	Scaled RMS residual
Matérn 3/2	$\ell = 0.35$	8.45×10^{-7}	2.29×10^{4}	3.194	0.0127
Matérn 3/2	$\ell = 0.50$	2.43×10^{-6}	7.93×10^{4}	1.591	0.0582
Matérn 3/2	$\ell = 0.75$	8.15×10^{-6}	3.96×10^{5}	1.277	0.1000
Matérn 3/2	$\ell = 1.00$	1.93×10^{-5}	1.36×10^{6}	1.087	0.1142
Matérn 3/2	$\ell = 1.50$	6.49×10^{-5}	7.67×10^{6}	1.258	0.1058
Matérn 3/2	$\ell = 2.00$	1.54×10^{-4}	2.45×10^{7}	1.276	0.0978
Matérn 3/2	$\ell = 3.00$	5.18×10^{-4}	1.13×10^{8}	1.316	0.0885
Cubic polyharmonic	$\phi(r) = r^3$	1.82×10^{-8}	1.63×10^{9}	1.385	0.0956

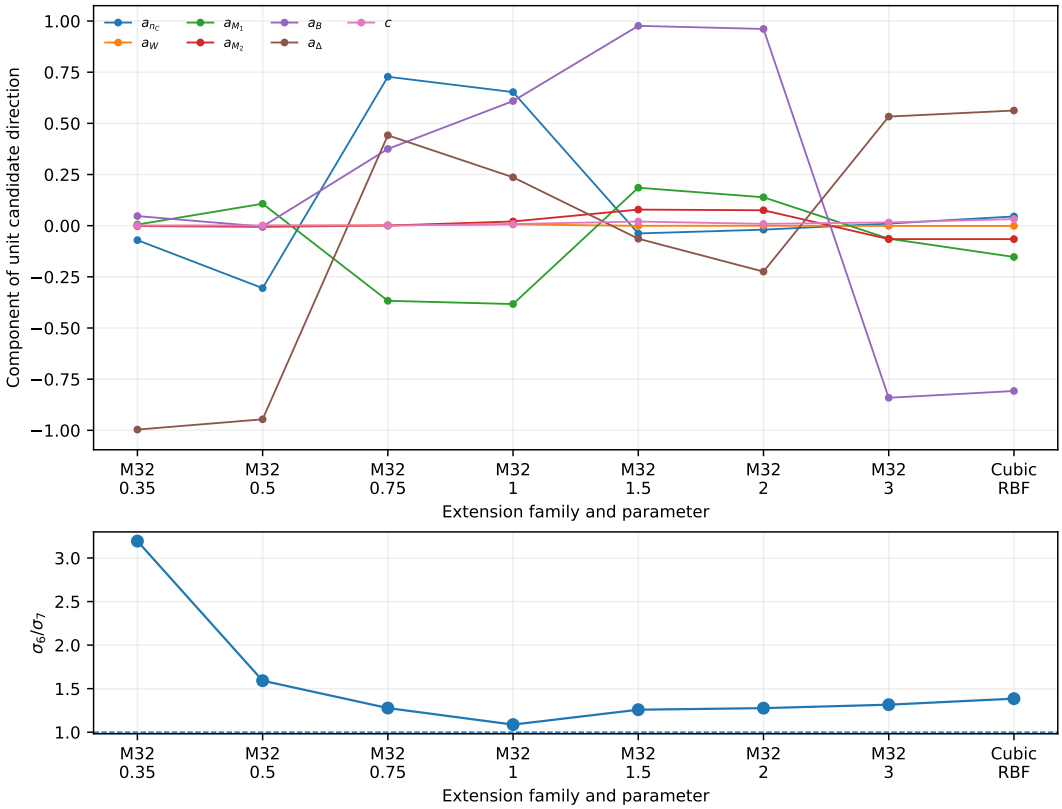


Fig. 6 Sensitivity of the least-squares PDE coefficient direction to the continuous-extension choice. The seven Matérn 3/2 length scales are shown alongside one cubic polyharmonic radial-basis interpolant. The upper panel gives the components of the unit smallest-singular-vector direction in standardized descriptor coordinates; the lower panel gives the ratio of the two smallest singular values

The maximum absolute difference between analytic gradients and their central finite-difference checks was 2.77×10^{-4} in standardized coordinates across the eight fits. This numerical check supports the derivative implementation, but it does not remove the ill-conditioning of the larger-length-scale Matérn kernel matrices or the augmented cubic interpolation system. The extension experiment is therefore a controlled sensitivity test, not evidence that the chosen Matérn family is the uniquely appropriate representation of molecular descriptor space.

4 Discussion

The results distinguish three questions that should not be combined. The descriptor-collision check asks whether the selected inputs can define an exact single-valued response. The finite-site theorem asks whether the observed values determine derivatives or sampled PDE residuals. The rank test asks whether coefficients are unique after a derivative field has been selected. Each check addresses a different source of ambiguity. Passing one does not imply that the others have been passed.

The residual-controllability theorem is particularly relevant to data-driven equation discovery. If gradients are not independently constrained, an interpolant can be constructed to make a chosen first-order equation hold at the finite observations, or to give any prescribed residuals there. A low residual on the same finite sites is therefore not sufficient evidence for a governing equation. PDE discovery methods based on resolved space–time data address a different information setting, in which measurements are collected over a field and derivative structure can be estimated from the observations [8, 9]. The present framework does not replace those methods; it identifies a preliminary admissibility issue when the inputs are isolated molecular descriptor vectors.

The six-descriptor model predicts boiling points better than carbon number alone within the 150-alkane dataset, and its advantage remains under the more demanding LOCO split. However, the data are concentrated in C8–C10, and the singleton classes are poorly represented. The 52.57 °C error for the held-out C1 class is a practical warning against relying on pooled performance alone. The six descriptors also produce two conflicting pairs of inputs. A richer graph representation may distinguish those structures, but that would be a QSPR improvement rather than evidence of a PDE.

The literature already contains regression studies using Zagreb-vector and VL descriptors for aliphatic-hydrocarbon properties, methods designed to address descriptor degeneracy in QSPR, and broader graph-descriptor and machine-learning approaches to property prediction [3–6, 15, 17]. The present results are not intended to outperform those larger models. The benchmark is used to show that ordinary QSPR prediction remains possible even when the conditions for PDE identification are not met.

The underlying ingredients of this paper are not claimed as new in isolation. Smooth extension from prescribed data has a long mathematical history, and rank-based methods are standard tools in parameter-identifiability analysis [10, 11, 14]. The paper's contribution is the way these ingredients are assembled and tested for a specific claim that arises in molecular QSPR: whether a finite table of graph descriptors and property values can support a first-order descriptor-space PDE. This is distinct from the descriptor-engineering response to degeneracy developed in graph-signal-processing QSPR, which aims to create more discriminating molecular features [17]. The proposed test instead combines a descriptor-consistency obstruction, a finite-site residual-controllability result, conditional coefficient-rank analysis, characteristic-data requirements and a numerical comparison of Matérn and cubic radial-basis extensions.

A targeted search of the literature located established PDE-discovery methods, general PDE-identifiability work, graph-based QSPR studies and 2026 regression work using topological indices for alkane properties. It did not locate a study that reports this complete sequence as an admissibility test for molecular descriptor tables. This is a bounded literature-search statement,

not proof that no related work exists. A broader database search and expert review may identify additional related studies.

The University of Pisa dataset page describes 150 acyclic alkanes and 1331 graph vertices [22]. The original raw graph strings are not included in the archive because the landing page does not state an explicit redistribution licence. Before the supplementary data are publicly deposited, the authors should confirm the terms with the dataset provider. The original 11-compound table is kept separate from the 150-molecule benchmark and is not used to fit the reported models.

5 Limitations

The conclusions are limited by the available data and by the mathematical class considered. The alkane benchmark contains only acyclic hydrocarbons with at most ten carbon atoms and is strongly imbalanced by carbon number. It does not support conclusions about cyclic, aromatic, heteroatom-containing or longer-chain molecules. The six graph descriptors are incomplete structural summaries and generate conflicting descriptor collisions. The NIST audit of the original 11-compound table is selective; a complete provenance-based review of every source value remains necessary. The MV1 and VL columns in that table could not be reproduced because their definitions were not supplied.

The numerical extension experiment uses a Matérn 3/2 family at seven length scales and one cubic polyharmonic radial basis extension. The coefficient directions differ across these selected extensions, but other extension families or descriptor scalings may produce other directions. The collision-only RMSE lower bound is a mathematical consequence of the duplicated descriptor vectors and does not imply that the regression model's overall error is dominated by the collisions. The kernel matrices become increasingly ill-conditioned at larger length scales. The rank analysis concerns first-order, constant-coefficient, homogeneous equations. Variable-coefficient, nonlinear or higher-order equations require different design matrices and corresponding identifiability analyses. The theorem assumes distinct descriptor sites; repeated sites with conflicting responses must be handled first by a declared statistical or aggregation model.

Finally, the study identifies no governing PDE for alkane boiling points. Its result is a framework for testing whether such a claim is supported by a particular dataset and set of assumptions. A broader predictive claim would require an independently curated external dataset, improved descriptor representations and validation across additional chemical families.

6 Conclusion

A finite molecular descriptor table can support a QSPR regression model, but the table alone does not determine the descriptor gradients required for a PDE. The three-gate test developed here begins by checking whether the chosen descriptors define a single-valued response. It then uses a finite-site residual-controllability theorem to show that observed values do not determine sampled first derivatives. Finally, it tests coefficient uniqueness only after a continuous extension and derivative field have been fixed, and it requires suitable characteristic data to determine the solution. In the 150-alkane benchmark, the six-descriptor representation contains two conflicting descriptor collisions. A nested grouped ridge model gives an MAE of 4.09 ± 0.14 °C under repeated grouped five-fold validation and a pooled MAE of 5.67 °C when a carbon-number class is held out. The extension experiment gives candidate coefficient directions with pairwise angles up to 88.03° across the seven Matérn length scales and one cubic polyharmonic extension. These results show that prediction and differential-law identification are separate tasks.

Funding. No specific funding was received for this work.

Consent to participate. Not applicable.

Consent for publication. Not applicable.

Data availability. The University of Pisa Alkanes Dataset analysed in this study is available from the Computational Intelligence and Machine Learning Laboratory, University of Pisa, Italy, at https://ciml.di.unipi.it/software/dataset/alkanes-dataset. The original molecular graph encodings have not been redistributed by the author. The derived alkane descriptor table and molecule-level prediction files are not included in the public data archive. These data can be regenerated locally using the provided scripts and a separately obtained copy of the original dataset, subject to the provider's applicable reuse conditions. The synthetic PDE grid and associated residual data are available at Figshare: https://doi.org/10.6084/m9.figshare.34944661.

Code availability. The Python scripts developed for descriptor construction, QSPR validation, and numerical extension-sensitivity analysis are publicly available through Figshare under the GNU General Public License version 3 or later (GPL-3.0-or-later): https://doi.org/10.6084/m9.figshare.34944649. The archive does not contain the original molecular graph encodings or the derived alkane descriptor table. Reproduction of the alkane analysis requires the original dataset to be obtained separately from the University of Pisa.

Author contributions. Syed Eirfan Atthar: Conceptualization, methodology, formal analysis, investigation, software, data curation, validation, visualization, writing: original draft, and writing: review and editing. The author approved the final manuscript.

Declarations

Competing interests. The author declares no competing financial or non-financial interests.

Ethics approval. Not applicable. This study uses previously available molecular graph and physicochemical-property data and includes a synthetic numerical example. No human participants or animals were involved.

References

[1] H. Wiener, Structural determination of paraffin boiling points, *J. Am. Chem. Soc.* **69**(1), 17–20 (1947). https://doi.org/10.1021/ja01193a005.

[2] A. Micheli, Neural network for graphs: a contextual constructive approach, *IEEE Trans. Neural Netw.* **20**(3), 498–511 (2009). https://doi.org/10.1109/TNN.2008.2010350.

[3] A. R. Nizami, S. F. Ali, M. Z. Afzal, and M. Inc, Novel descriptors for the prediction of molecular properties, *Open Chem.* **23**(1), 20250194 (2025). https://doi.org/10.1515/chem-2025-0194.

[4] R. Bounaceur, F. Paes, R. Privat, and J.-N. Jaubert, AI-powered prediction of critical properties and boiling points: a hybrid ensemble learning and QSPR approach, *J. Cheminform.* **17**, 132 (2025). https://doi.org/10.1186/s13321-025-01062-9.

[5] J. W. Burns and W. H. Green, Generalizable, fast, and accurate DeepQSPR with fastprop, *J. Cheminform.* **17**, 73 (2025). https://doi.org/10.1186/s13321-025-01013-4.

[6] H. M. Fraz, Z. ul Haque, K. Ali, M. F. Nadeem, P. Roy, A. Li et al., A computational QSPR study of neurotransmitters using distance-based topological descriptors, *Chem. Pap.* (2026). https://doi.org/10.1007/s11696-026-05962-4.

[7] S. L. Brunton, J. L. Proctor, and J. N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems, *Proc. Natl. Acad. Sci. USA* **113**(15), 3932–3937 (2016). https://doi.org/10.1073/pnas.1517384113.

[8] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz, Data-driven discovery of partial differential equations, *Sci. Adv.* **3**(4), e1602614 (2017). https://doi.org/10.1126/sciadv.1602614.

[9] S. L. Brunton and J. N. Kutz, Promising directions of machine learning for partial differential equations, *Nat. Comput. Sci.* **4**, 483–494 (2024). https://doi.org/10.1038/s43588-024-00643-2.

[10] Y. Salmaniw and A. P. Browning, Structural identifiability of linear-in-parameter parabolic PDEs through auxiliary elliptic operators, *J. Math. Biol.* **91**, 4 (2025). https://doi.org/10.1007/s00285-025-02225-w.

[11] M. J. Simpson and R. E. Baker, Parameter identifiability, parameter estimation, and model prediction for differential equation models, *SIAM Rev.* **68**(1), 153–171 (2026). https://doi.org/10.1137/24M1667968.

[12] Y. He, H. Zhao, and Y. Zhong, How much can one learn a partial differential equation from its solution? *Found. Comput. Math.* **24**(5), 1595–1641 (2024). https://doi.org/10.1007/s10208-023-09620-z.

[13] Y. S. Ju, Structural identifiability of constitutive relations from indirect field observations, *PNAS Nexus* **5**(9), pgag281 (2026). https://doi.org/10.1093/pnasnexus/pgag281.

[14] H. Whitney, Analytic extensions of differentiable functions defined in closed sets, *Trans. Amer. Math. Soc.* **36**(1), 63–89 (1934). https://doi.org/10.2307/1989708.

[15] V. H. Narendra and A. G. Puranik, Multiple linear regression-based predictions of physicochemical properties of aliphatic hydrocarbons using Zagreb vector index and VL-index, *Asian J. Math. Comput. Res.* **33**(2), 72–80 (2026). https://doi.org/10.56557/ajomcor/2026/v33i210576.

[16] V. H. Narendra and P. Mahalakshmi, Computation of Zagreb vector index and total eccentricity in the physical properties of linear alkanes, *Asian J. Phys. Chem. Sci.* **14**(1), 11–17 (2026). https://doi.org/10.9734/ajopacs/2026/v14i1277.

[17] X. Song, G. Wen, and L. Chai, Graph signal processing based nonlinear QSAR/QSPR model learning for compounds, *Biomed. Signal Process. Control* **91**, 106011 (2024). https://doi.org/10.1016/j.bspc.2024.106011.

[18] D. R. Burgess, Jr. and A. P. Hamins, *Heats of Combustion and Related Properties of Pure Substances*, NIST Technical Note 2126-upd1 (National Institute of Standards and Technology, Gaithersburg, MD, 2026). https://doi.org/10.6028/NIST.TN.2126-upd1.

[19] National Institute of Standards and Technology, NIST Chemistry WebBook, SRD 69: Propane, phase-change data (accessed 7 November 2026). https://webbook.nist.gov/cgi/cbook.cgi?ID=C74986&Mask=1ACCE.

[20] National Institute of Standards and Technology, NIST Chemistry WebBook, SRD 69: Butane, normal melting-point data (accessed 7 November 2026). https://webbook.nist.gov/cgi/cbook.cgi?ID=C106978&Type=TFREEZE.

[21] National Institute of Standards and Technology, NIST Chemistry WebBook, SRD 69: Undecane, phase-change data (accessed 7 November 2026). https://webbook.nist.gov/cgi/cbook.cgi?ID=C1120214&Mask=C&Units=SI.

[22] Computational Intelligence and Machine Learning, University of Pisa, Alkanes Dataset (150 acyclic alkanes; accessed 7 November 2026). https://ciml.di.unipi.it/software/dataset/alkanes-dataset.

[23] A. Micheli, The alkanes data set: description and target-value scaling (University of Pisa, 2009). https://pages.di.unipi.it/micheli/dataset/alkanes-description.txt.

[24] A. E. Hoerl and R. W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, *Technometrics* **12**(1), 55–67 (1970). https://doi.org/10.1080/00401706.1970.10488634.

[25] M. L. Stein, *Interpolation of Spatial Data: Some Theory for Kriging* (Springer, New York, 1999). https://doi.org/10.1007/978-1-4612-1494-6.

[26] H. Wendland, *Scattered Data Approximation*, Cambridge Monographs on Applied and Computational Mathematics, Vol. 17 (Cambridge University Press, Cambridge, 2005).