CoolData: Benchmarking Machine Learning Methods for Electronics Cooling

Anonymous authors
Paper under double-blind review

Abstract

The integration of Machine Learning (ML) into Computer-Aided Engineering (CAE) workflows promises significant speed-ups, particularly in the field of Computational Fluid Dynamics (CFD). However, the development of robust ML models in CFD is hindered by the lack of high-quality, application-specific datasets. This work addresses this gap with CoolData, a large-scale electronics cooling dataset containing 60,848 stationary 3D flow and temperature fields for a diverse set of geometries, simulated using the industrial CFD solver Simcenter STAR-CCM+. To our knowledge, this is the first high-quality and publicly available dataset for electronics cooling applications.
We show how this dataset can be leveraged in an ML pipeline, benchmarking both volumetric models (3D U-Net, DOMINO, Transolver) and surface models (MeshGraphNet, DOMINO, Transolver) across a range of training-set sizes. Beyond in-distribution accuracy, we introduce a systematic out-of-distribution (OOD) study along two axes enabled by CoolData's controlled parametrization: geometric composition (body count and shape held out from training) and thermal extrapolation (peak component temperature beyond the training range).

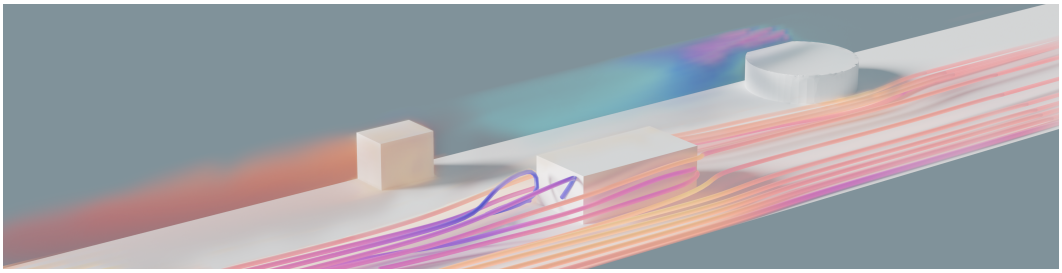


Figure 1: Visualization of multiple data fields of a single sample from our dataset. The figure illustrates the temperature field (volume rendering, left), the velocity field (streamlines), and the turbulent kinetic energy (volume rendering, right).

1 Introduction

Following advances in text, image, and video generation, Machine Learning (ML) is rapidly progressing toward physics and engineering intelligence, it is viewed as a core enabler to overcome critical computational bottlenecks in engineering workflows (1),in particular accelerating Computational Fluid Dynamics (CFD). CFD represents one of the most resource-intensive steps in modern engineering design, often involving billions of variables and requiring compute times of days to weeks (2). The complexity, time, and effort required to generate accurate predictions slow down simulation adoption and limit overall engineering efficiency. A broad spectrum of ML architectures has been proposed to address this challenge, including Neural Operator (3), Transformer-based (4; 5), and iterative solver-like architectures (6), and, most recently, Engineering Foundation Models (CAE-FMs).

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.

The central bottleneck, however, is data availability (7). Unlike text or image data, engineering simulation data is scarce due to Intellectual Property (IP) concerns and efforts required for (synthetically) generating, storing, and processing it (8; 7). To put this in perspective: while the original OpenAI Codex model was trained on approximately 159 GB of Python code extracted from over 54 million GitHub repositories (9), a single high-fidelity CFD simulation can easily exceed 100 GB. Beyond sheer volume, existing public datasets are further limited in application diversity, predominantly covering simplified external aerodynamics geometries with only modest variation around selected base shapes (10; 11; 12; 13; 14). While physical phenomena often admit low-dimensional approximations (15), geometric design spaces are potentially unbounded, limiting model generalization and complicating scaling studies.

Electronics cooling is essential in electrical and electronic systems, where thermal constraints often limit performance (16). With often highly constrained geometrical conditions on cooling, with flow patterns dominated by wall-bounded turbulence, heat recirculation, and complex conduction paths is specifically challenging (17; 18). CFD simulations are the standard tool for this purpose, solving the Navier-Stokes equations with turbulence models such as RANS (19). Despite their physical fidelity, such simulations remain computationally expensive, making fast surrogate models highly desirable.

We introduce **CoolData**, a large-scale, high-fidelity CFD dataset addressing electronics cooling applications. CoolData comprises over $60,000$ stationary 3D thermal-fluid simulations for parametrically varied, geometrically complex configurations. The data is generated with the industrial solver Simcenter STAR-CCM+ (20). To generate the data, we adopt a controlled geometry variation strategy based on up to 35 distinct parameters. This serves a dual purpose: it captures configurations typical of electronics cooling practice while enabling rigorous studies of data scaling behaviour, making CoolData a valuable resource for investigating the data requirements of ML and CAE-FM. The dataset is published on Hugging Face (`https://anonymous-hf.com/a/grymmk4n2b1c/`) under a CC BY-NC 4.0 license, with an accompanying accessor library.

Our main contributions are:

- **A large-scale, open CFD dataset** of $60,848$ high-fidelity 3D thermal-fluid simulations for geometrically diverse electronics cooling configurations ($\sim 3\,\text{TB}$ data in total).
- **A controlled geometry variation strategy** enabling systematic studies of data scaling behaviour of ML in an industrially relevant setting.
- **ML benchmarks** evaluating model accuracy as a function of data availability, comparing four architectures across two benchmarks with a focus on in-distribution as well as out-of-distribution benchmarking.
- **An accessor library** providing tools for loading, visualizing, and converting the dataset into suitable formats.

2 Related Work

CFD Datasets for ML: Several open datasets have been released to support ML for CFD, spanning a wide range of physics and geometric complexity. Academic benchmarks have traditionally relied on canonical geometries such as airfoils, bluff bodies, or duct flows, which are well understood and easy to validate (21; 22; 23). Recently, multi-physics and multi-domain benchmarks such as EAGLE (24) and The Well (8) have broadened the scope of available training data on academic examples significantly. However, in many cases these are limited to very simple geometries. More recently, large-scale industrial-grade datasets have emerged, particularly in external aerodynamics: the Windsor body (25) and AhmedML (14) datasets offer high-fidelity CFD results for simplified road vehicle geometries; DrivAerML (12) and DrivAerNet++ (13) extend this to more realistic car shapes; and HiLiftAeroML (26) targets high-lift aircraft configurations using wall-modeled LES. Despite this progress, existing datasets focus almost exclusively on external aerodynamics or on academic flow configurations, with limited geometric variation.

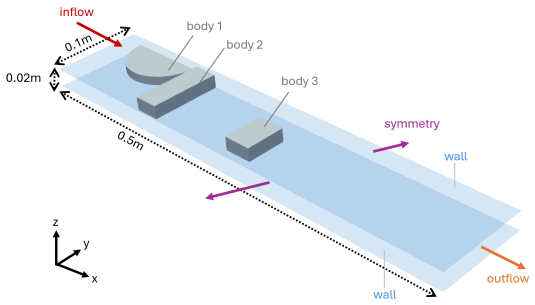


Figure 2: Dimensions and structure of the computational domain with 3 exemplary bodies.

ML for Electronics Cooling: ML has shown increasing promise for thermal management applications, with studies demonstrating its utility for surrogate modeling of microchannel heat sinks, thermoelectric coolers, and phase-change material configurations (27; 28). For example, Sikirica et al. (27) leveraged ML to predict thermal resistance and pumping power in microchannel heat sinks, significantly reducing computational cost compared to full CFD evaluations. However, industrially relevant datasets in this space with sufficient geometric diversity are scarce. Existing work predominantly relies on simplified geometries (such as rectangular channels, pin-fin arrays, or duct flows), consequently, they are not suited to study ML behaviour in scenarios where specifically geometries are far more varied.

Gap Addressed by CoolData: Most existing efforts in this space are constrained by limited geometry variation, small scale, or restricted access. CoolData addresses this by providing a large, publicly available dataset with systematic geometric variation, enabling both surrogate-model training and rigorous data-availability benchmarking.

3 Dataset

Baseline Geometry and Flow Conditions

This dataset focuses on (power-) electronics applications, comprising typical electronic components (such as chips, capacitors, coils, and transistors) mounted on Printed Circuit Boards (PCBs) and enclosed within box-like housings. Thermal management is achieved through convective air cooling, a common and practically relevant cooling strategy. The specific configuration considered in the contribution is illustrated in Figure 2.Air enters the electronics compartment from the left (at $x = 0$) and exits on the right, modelled by a constant, parametrized velocity and a fixed temperature of $T_{\text{in}} = 20\,^{\circ}\text{C}$ at the inlet, and a fixed pressure of zero at the outlet. The top and bottom boundaries are walls with a no-slip boundary condition with a wall model, whereas the right and left have symmetry boundary conditions, enabling the modelling of a much larger, symmetrical domain than the size of the computational domain, which is more realistic for electronics cooling scenarios. Electronic components are geometrically abstracted as simple cuboids or cylinders mounted on the bottom wall. To promote a numerically stable and well-developed outflow, all components are positioned toward the inlet side of the domain, away from the outlet. We focus here on the air-flow (which is computationally the most demanding aspect), i.e., prescribing the body temperatures and not modelling the internal heat conduction. This appropriately simplifies the computation and the setup, while not affecting the realism of the simulation. Heat conductivity in the corresponding chips would be rather high such that temperature variations are rather small, compared to overall temperature differences. Thus, sampling different body temperatures is equivalent to sampling different heat loading conditions.

Parametrization and Generation of Geometry Variants

The domain for all samples in the dataset is a channel of fixed size as introduced above. Inside the domain, there are between 2 and 6 heated, wall-mounted bodies, either in a cuboid

or cylindrical shape. While cuboids are often used to simulate PCB components, since they generate most of the flow features exhibited in PCB cooling, the addition of cylinders allows for the approximation of other components like capacitors and helps generalization. The bodies' positions, sizes, and independent temperature are parametrized. This results in a total variation of a maximum of 35 parameters (depending on the number of involved geometries) for geometry, heating, and inflow (cf., Table 7 in Appendix B).

Parameter combinations are generated using Latin Hypercube Sampling (29) with equal probabilities. Each configuration contains two mandatory cuboids and up to two additional cuboids and two cylinders, each included with probability of 80%. All parameters are sampled independently. Bodies extending beyond the lateral domain boundaries are clipped, consistent with the symmetric setup. For overlapping bodies, only fluid-exposed faces are created, avoiding conflicting boundary conditions for fully enclosed bodies. For partially overlapping bodies, each exposed face retains its body's independently sampled temperature T_{Body}, resulting in piecewise-constant Dirichlet conditions on composite surfaces.

Although PCB and housing configurations span an infinite-dimensional space, we adopt an idealized setup with 35 latent dimensions, following standard external aerodynamic datasets (12). This controlled setting enables efficient sampling and systematic studies of data requirements and out-of-distribution prediction, while capturing practical challenges such as irregular component arrangements and narrow channels between closely spaced components.

CFD Workflow

Adopting standard industrial practice and tools, subject to extensive (experimental) validation (e.g., (30)), we consider CFD results as ground truth and do not consider further experimental evidences. For all geometries we use the same CFD model following industrial modelling practices. That is, we employ a steady state Reynolds-averaged Navier–Stokes equations (RANS) model with a $k-\varepsilon$ turbulence model and an appropriate wall layer model to appropriately reflect the fluid flow. The thermal behaviour is modeled by means of a segregated fluid energy model. The model is implemented in Simcenter STAR-CCM+ and the corresponding `.sim`-file is provided along the dataset. For further details, such as choice of specific solver parameters we refer to the provided `.sim`-file in the dataset and documentation in the user manual (20).

Meshing: We use a trimmed hexahedral cell mesher (base size 2 mm, 3 prism layers near boundaries) with typically 1.6×10^5 cells per geometry. A mesh sensitivity study on 3 representative designs confirmed the chosen resolution is sufficient (coefficients of variation below 10^{-5} for key quantities of interest); full parameters are in the provided `.sim` file.

Convergence Criteria: Our dataset focuses on steady flows. When the main quantities of interest, i.e., average velocity, average temperature, and heat transfer at the outflow boundary, exhibit asymptotic behaviour the corresponding simulation is considered to be converged. We use $|\max - \min|$ over the last 30 iterations as a criterion, i.e., stop the simulation when these have reached 1.0×10^{-4}, 1.0×10^{-5}, and 1.0×10^{-4}, respectively. Residuals are expected to decrease during convergence but may vary in magnitude across designs. To address this, a standard deviation criterion of 0.01 over the last 100 iterations ensures residuals stabilize before stopping the simulation. Furthermore, we filter out designs with large, oscillatory residuals characteristic of transient flows which have not converged after $1,500$ iterations[1].

Dataset Contents and Availability

All data points in the dataset have a unique ID, ranging from 1 to 60,848, which is used to identify the corresponding file names (`volume_design_<ID>_p.cgns` and `surface_design_<ID>_p.cgns`). For sampling efficiency, we excluded simulations that had not converged after 1,500 iterations (cf., Section Convergence Criteria); corresponding

[1]Even using a RANS model, mathematically, a steady flow cannot be guaranteed for all cases.

cases are typically edge cases, such as extremely narrow channels. Furthermore, the available compute budget was oversubscribed, such that not all sampled configurations were simulated. The dataset consists of 60,848 data points, with a total compressed size of approximately 1 TB (995 GB as distributed on Hugging Face), corresponding to an uncompressed size of approximately 3 TB.

For each data point we provide geometry parameters (`metadata.parquet`), volume field quantities (velocity, pressure, temperature, turbulent kinetic energy and its dissipation rate), and wall-related surface quantities (area, normal, temperature, heat transfer coefficient, pressure, wall shear stress), see Appendix I for the complete field list with units.

The dataset provides a comprehensive set of typical quantities commonly used in CFD applications such as flow prediction or heat transfer modeling (cf., Appendix I). Specifically they allow the calculation of further derived quantities (e.g., pressure coefficient, total pressure coefficient, skin friction coefficient, temperature coefficient, or heat flux) or integral quantities (e.g., total pressure drop, component junction temperatures, thermal resistance, or fan operating point parameters). An analysis of selected thermal performance metrics across the dataset is given in Appendix E.

4 Benchmarks

Given the size of the dataset, it provides a unique opportunity to benchmark different models not only in terms of their performance but also in terms of their data requirements. Within this contribution we benchmark 4 representative ML models: a volumetric model (3D U-Net), a surface-centric model (MeshGraphNet), and two models evaluated on both volumetric and surface-centric targets (Transolver, DOMINO). We thereby focus on direct field prediction rather than prediction of derived or integrated quantities. The latter can be determined using appropriate post-processing from the predicted field quantities (cf., Appendix I). Our volumetric benchmarks target the temperature field, while our surface-centric benchmarks target the wall shear stress magnitude, both being primary quantities of interest for electronics cooling design. Wall shear stress is particularly central here, as it is closely linked to the local convective heat transfer coefficient, the key metric governing cooling performance.

ML-based field prediction from 3D meshes using geometric deep learning

As a first benchmark we evaluate the capability of ML models to predict physics field quantities, along two complementary axes: a volumetric benchmark, comparing 3D U-Net (31), Transolver (32), and DOMINO (6) on the temperature field, and a surface-centric benchmark, comparing MeshGraphNet (33), Transolver, and DOMINO on the wall shear stress magnitude. Each benchmark evaluates 3 models; since Transolver and DOMINO appear in both, this amounts to 4 distinct architectures in total. 3D U-Net is implemented via `pytorch-3dunet` (34), whereas MeshGraphNet, Transolver, and DOMINO are implemented via NVIDIA PhysicsNeMo (35). For more details on the specific training we refer to Appendix F. The obtained performance metrics for the surface-centric (wall shear stress) benchmark are shown in Table 2, and for the volumetric (temperature) benchmark in Table 1; selected results are shown in Figure 3. For both, results have been obtained using $50,000$ samples over 100 epochs on 8×NVIDIA H100 GPUs . Inference times are reported per sample as the average over 100 runs on 4×NVIDIA L4 GPUs (all models within each benchmark evaluated on identical hardware to ensure comparability).

Table 1: Comparative analysis of deep learning models for the volumetric (temperature) benchmark using 50,000 training samples. Error is evaluated on 1,000 in-distribution test samples, pooled over all points rather than averaged per-sample (see Appendix G).

Model	MSE	MAE	Max AE	$\mathbf{R^2}$	$\mathbf{T}_{\text{train}}$	$\mathbf{T}_{\text{infer}}$	# Params
3D U-Net (31)	19.25868	1.977 K	64.026 K	0.8595	**1 h 30 m**	**47 ms**	1,184,978
Transolver (32)	**4.97988**	**1.160 K**	**56.757 K**	**0.9046**	17 h 21 m	266 ms	1,151,537
DOMINO (6)	5.53310	1.281 K	60.709 K	0.8940	8 h 33 m	110 ms	1,196,807

The results in Table 1 highlight a clear hierarchy in predictive performance across the 3 volumetric architectures. Transolver achieves the best result on every accuracy metric (MSE, MAE, Max AE, R^2), with MAE=1.16 K and R^2=0.9046. DOMINO follows closely, within 0.10 K MAE and 0.011 R^2 of Transolver, despite its markedly different decomposable multi-scale architecture. 3D U-Net trails both in accuracy (MAE=1.98 K, R^2=0.8595), consistent with a resolution ceiling imposed by its fixed voxel grid, which caps the geometric detail available to the model regardless of training set size but is by far the fastest of the three at inference (47 ms vs. 266 ms for Transolver and 110 ms for DOMINO).

Table 2: Comparative analysis of ML models for the surface-centric (wall shear stress) benchmark using 50,000 training samples. Error is evaluated on 1,000 in-distribution test samples, pooled over all points rather than averaged per-sample (see Appendix G).

Model	**MSE**	**MAE**	**Max AE**	$\mathbf{R^2}$	$\mathbf{T_{train}}$	$\mathbf{T_{infer}}$	**# Params**
MeshGraphNet (33)	0.19742	0.127 Pa	16.886 Pa	0.6452	3h 20m	11.648 ms	1,295,489
Transolver (32)	0.09348	**0.078 Pa**	14.834 Pa	0.8320	5h 40m	**5.163 ms**	1,153,073
DOMINO (6)	**0.08329**	0.099 Pa	**14.739 Pa**	**0.8503**	**2h 10m**	16.560 ms	1,144,475

The results in Table 2 show a less clear-cut hierarchy for the surface-centric (wall shear stress) benchmark than for the volumetric case below. DOMINO achieves the best MSE (0.083), Max AE (14.74 Pa), and R^2 (0.8503), while Transolver achieves the lowest MAE (0.078 Pa) and is by far the fastest of the three at inference (5.2 ms vs. 16.6 ms for DOMINO and 11.6 ms for MeshGraphNet), notably the reverse of the volumetric benchmark's speed ordering, where Transolver was the slowest of the three models evaluated there. MeshGraphNet trails both dual-role architectures on every accuracy metric (MAE=0.143 Pa, R^2=0.6452).

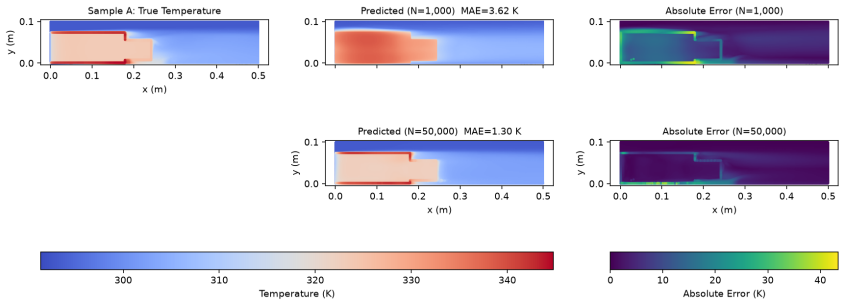


Figure 3: Comparison of true (CFD ground truth) and Transolver-predicted volumetric temperature fields for a representative test sample, evaluated at two training set sizes (1000 and 50000). More sample visualizations are available in Appendix21

Out-of-distribution (OOD) generalization

CoolData's controlled parametrization enables systematic OOD studies along many independent axes. Here, we focus on geometric composition and thermal extrapolation as representative studies, the former for both volumetric (temperature) and surface (wall shear stress) predictions. For the geometric axis, we denote each body composition as **n**q**m**c, i.e., **n** cuboids and **m** cylinders, and train exclusively on the 3q0c composition (three cuboids, no cylinder). We withhold all other body-count/shape combinations from training and evaluate generalization to configurations with fewer bodies (2q0c), altered shape at matched body count (2q1c, a cuboid replaced by a cylinder), and increased body count (4q0c, 4q2c). For the thermal axis, we cap the training distribution's peak active-body temperature at $T_{\text{cap}} = 60\,^{\circ}\text{C}$ and evaluate on three held-out bands: in-distribution ($T < 60\,^{\circ}\text{C}$), moderate extrapolation ($60\,^{\circ}\text{C} \le T < 72\,^{\circ}\text{C}$, i.e. $T_{\text{cap}} \le T < 1.2\,T_{\text{cap}}$), and severe extrapolation ($T \ge 72\,^{\circ}\text{C}$). T_{cap} is a threshold chosen for this study and is distinct from the dataset's own sampled temperature range (body temperature $T_{\text{Body}} \in [20\,^{\circ}\text{C}, 80\,^{\circ}\text{C}]$: even the severe-extrapolation band remains within CoolData's physically sampled range, so this study evaluates extrapolation beyond what a model was *permitted to see during training*, not beyond what the underlying data-generating process can produce.

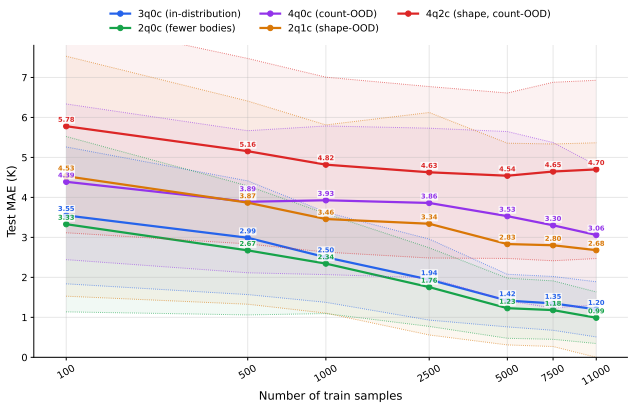


Figure 4: Body-composition OOD learning curves for Transolver on the volumetric benchmark: temperature test MAE (mean ± std, shaded) versus number of training samples, for five body compositions. 3q0c is in-distribution; all other compositions are held out entirely from training. Results for UNet3D+FiLM and DOMINO are shown in Appendix H. Exact values at $N_{\text{train}} = 11{,}000$ are reported in Table 3. For reference, the mean temperature rise above the inlet, averaged over all cells of 1,000 random samples, is 5.2 K.

Table 3: Body-composition OOD test MAE (K), mean ± std (sample-level dispersion across the held-out test set, not cross-seed variance), at $N_{\text{train}} = 11{,}000$. 2q0c/2q1c/4q0c/4q2c are held out entirely from training; 3q0c is in-distribution. Full training size ($N = 100$-11,000) shown in Figure 4 (Transolver) and Appendix H (UNet3D+FiLM, DOMINO).

Model	**3q0c** (in-dist.)	**2q1c** (shape-OOD)	**2q0c** (count-OOD)	**4q0c** (count-OOD)	**4q2c** (shape, count-OOD)
Transolver	$\mathbf{1.20 \pm 0.69}$	$\mathbf{2.68 \pm 2.69}$	$\mathbf{0.99 \pm 0.65}$	$\mathbf{3.06 \pm 1.75}$	4.70 ± 2.23
UNet3D+FiLM	2.29 ± 1.05	3.84 ± 2.43	1.71 ± 0.95	3.19 ± 1.45	6.12 ± 2.73
DOMINO	1.74 ± 0.70	2.79 ± 2.33	1.48 ± 0.97	3.21 ± 2.18	$\mathbf{4.74 \pm 2.14}$

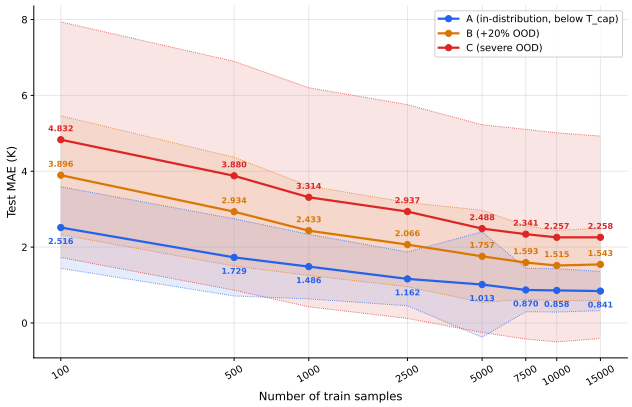


Figure 5: Temperature OOD learning curve for Transolver: test MAE (mean ± std, shaded) versus number of training samples, for in-distribution (A), moderate (B), and severe (C) thermal extrapolation. Exact values at $N_{\text{train}} = 15{,}000$ are reported in Table 4.

For the volumetric OOD benchmark and corresponding geometry variation, no single architecture dominates uniformly: DOMINO achieves the lowest mean error at the majority of (composition, N_{train}) combinations (20 of 35), typically at low-to-moderate training-set sizes, while Transolver's steeper learning curve lets it overtake DOMINO as N_{train} grows, most clearly on 3q0c, where Transolver leads from $N_{\text{train}} = 500$ onward and its gap over UNet3D+FiLM widens from 0.5 K at $N_{\text{train}} = 100$ to 1.1 K at $N_{\text{train}} = 11{,}000$ (Figure 4, Appendix H). UNet3D+FiLM does not achieve the lowest mean error on any composition

Table 4: Temperature OOD test MAE (K), mean ± std (sample-level dispersion across the held-out test set, not cross-seed variance), for Transolver at $N_{\text{train}} = 15{,}000$. A = in-distribution (peak active-body temperature $< T_{\text{cap}}$); B = moderate extrapolation ($T_{\text{cap}} \leq T < 1.2\,T_{\text{cap}}$); C = severe extrapolation ($T \geq 1.2\,T_{\text{cap}}$). Full training size ($N = 100$-$15{,}000$) shown in Figure 5.

Model	**A (in-distribution)**	**B (+20% OOD)**	**C (severe OOD)**
Transolver	0.84 ± 0.52	1.54 ± 0.96	2.26 ± 2.67

or training size in this test. For all three architectures, the error on 2q0c is slightly below the in-distribution error, whereas all compositions with additional bodies or a cylinder show higher errors (Table 3).

On the thermal OOD benchmark, we focus on the best performing model. Transolver's error at $N_{\text{train}} = 15{,}000$ rises from 0.84 ± 0.52 K in-distribution to 1.54 ± 0.96 K under moderate extrapolation and 2.26 ± 2.67 K under severe extrapolation (Table 4), i.e., the error increases gradually with the distance from the training range.

For the surface benchmark, we repeat the body-composition study for Transolver on the wall shear stress (Figure 6, Table 5). Here, removing a body (2q0c) or adding a cuboid (4q0c) leaves the error at the in-distribution level, whereas the compositions containing a cylinder show errors about 3 (2q1c) and 5 (4q2c) times higher at $N_{\text{train}} = 11{,}000$.

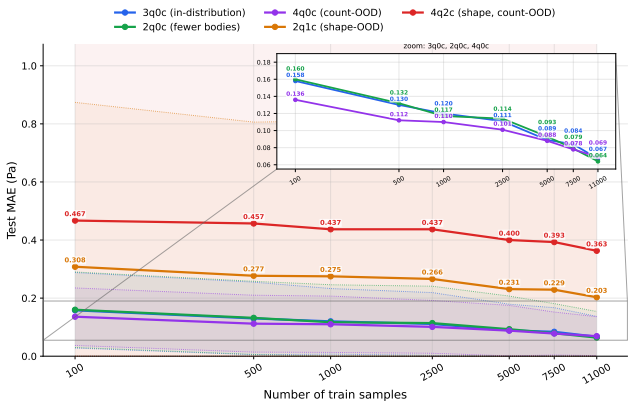


Figure 6: Body-composition OOD learning curves for Transolver on the surface benchmark: wall shear stress test MAE (mean ± std, shaded; clipped at zero and at the upper axis limit) versus number of training samples, evaluated on 90 test samples per composition. 3q0c is in-distribution; all other compositions are held out entirely from training. Inset: zoom on 3q0c, 2q0c, and 4q0c, whose curves nearly coincide. Exact values at $N_{\text{train}} = 11{,}000$ are reported in Table 5. For reference, the mean wall shear stress magnitude on the body surfaces is 0.24 Pa (3q0c), 0.23 Pa (2q0c), 0.25 Pa (4q0c), 0.46 Pa (2q1c), and 0.57 Pa (4q2c).

Table 5: Body-composition OOD wall shear stress test MAE (Pa), mean ± std (sample-level dispersion across the held-out test set, not cross-seed variance), for Transolver at $N_{\text{train}} = 11{,}000$, evaluated on 90 test samples per composition. 2q0c/2q1c/4q0c/4q2c are held out entirely from training; 3q0c is in-distribution. Full training size ($N = 100$-$11{,}000$) shown in Figure 6, except for 4q0c, which is reported in this table only.

Model	**3q0c** (in-dist.)	**2q1c** (shape-OOD)	**2q0c** (count-OOD)	**4q0c** (count-OOD)	**4q2c** (shape, count-OOD)
Transolver	0.067 ± 0.070	0.203 ± 0.466	0.064 ± 0.090	0.069 ± 0.067	0.363 ± 0.877

Even in-distribution, wall shear stress prediction carries a relative error of 29.0%, rising to 63.6% under severe geometric OOD (Table 6), indicating that WSS is intrinsically harder to

Table 6: Transolver mean relative error (WSS MAE as a percentage of the mean WSS magnitude) at $N_{\text{train}} = 11{,}000$. Reference means: 0.2312/0.5678/0.2484/0.2590/0.5708 Pa for 3q0c/2q1c/2q0c/4q0c/4q2c respectively, corresponding relative-error tables for the temperature and thermal-extrapolation studies are reported in Appendix H, Table 10.

Metric	**3q0c**	**2q1c**	**2q0c**	**4q0c**	**4q2c**
MAE (Pa)	0.067	0.203	0.064	0.069	0.363
Relative error (%)	29.0	35.8	25.8	26.6	63.6

learn than temperature at comparable absolute-error levels, independent of extrapolation difficulty.

These results indicate an exponential relationship between geometric complexity and prediction difficulty, most pronounced for wall shear stress, where even in-distribution error is substantial in relative terms (Table 6). In follow-up studies, we plan to quantify these challenges further.

5 Conclusion

In this contribution, we present **CoolData**: a large-scale electronics cooling dataset containing over 60,000 stationary 3D flow fields for a diverse set of geometries. Additionally, we share two example ML workflows, one using voxelized volumetric data and the other using point-based/graph-based surface representations, to demonstrate how to use the dataset for training ML models.

CoolData is, to the best of our knowledge, the first publicly available large-scale dataset for electronics cooling generated with an industrial-grade CFD solver (Simcenter STAR-CCM+). It extends industrial-grade benchmarks beyond external aerodynamics to thermal management, a domain of growing relevance with the electrification of many industries, and offers a structural diversity of geometries that makes it a challenging benchmark for ML models.

6 Limitations and Future Work

We see two major limitations concerning the current dataset: many real-world applications are not necessarily "close" to the data distribution presented and more realistic simulation methods are available, yet infeasible for a dataset of this scale.

Despite its diversity (cf. Section 3), the dataset does not cover many practical electronics components and cooling scenarios. Therefore, a broader range of geometries is needed to realize CAE-FM. Nevertheless, the dataset provides a key stepping stone, enabling controlled in- and out-of-distribution scaling studies to inform data requirements.

Following industry practice, we use stationary RANS simulations for computational feasibility. Transient methods such as Large Eddy Simulation (LES) and Detached Eddy Simulation (DES) would improve realism and relax geometric constraints such as the long downstream region required to ensure a recirculation-free outlet. However, storing individual transient states for all geometries would increase data volume hundreds of times, making storage and distribution infeasible. Even if only averaged quantities were stored, these methods would remain orders of magnitude more computationally expensive

Benchmarking is limited to four standard ML models commonly used in Computer Aided Engineering (CAE), with three models per benchmark. Out-of-distribution evaluation covers only extrapolation within the dataset's design space: unseen body counts and compositions, and temperatures beyond the training range. Temperature extrapolation is evaluated only for Transolver. Future work will assess additional models and more challenging out-of-distribution scenarios.

AI DISCLOSURE

In this work, we used generative AI tools to create or modify scientific figures or images, create or edit software code, summarize and analyse existing literature, brainstorming, sourcing/searching for information, edit a research paper to improve readability, identify relevant literature, format references, suggest a structure for a research paper, propose a title or keywords for a research paper. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

[1] Jan Paul Stein and Alessandro Faure Ragani. On the brink of a revolution? Engineering simulation in the age of AI, 2025.

[2] Engineering.com. Research Report: The State of Simulation, Prototyping and Validation. Technical report, Engineering.com, April 2022.

[3] Maximilian Herde, Bogdan Raonić, Tobias Rohner, Roger Käppeli, Roberto Molinaro, Emmanuel de Bézenac, and Siddhartha Mishra. Poseidon: Efficient Foundation Models for PDEs, May 2024.

[4] Benedikt Alkin, Andreas Fürst, Simon Schmid, Lukas Gruber, Markus Holzleitner, and Johannes Brandstetter. Universal physics transformers: A framework for efficiently scaling neural operators. *Advances in Neural Information Processing Systems*, 37:25152–25194, 2024.

[5] Huakun Luo, Haixu Wu, Hang Zhou, Lanxiang Xing, Yichen Di, Jianmin Wang, and Mingsheng Long. Transolver++: An accurate neural solver for PDEs on million-scale geometries. *arXiv preprint arXiv:2502.02414*, 2025.

[6] Rishikesh Ranade, Mohammad Amin Nabian, Kaustubh Tangsali, Alexey Kamenev, Oliver Hennigh, Ram Cherukuri, and Sanjay Choudhry. Domino: A decomposable multi-scale iterative neural operator for modeling large scale engineering simulations. *arXiv preprint arXiv:2501.13350*, 2025.

[7] Neil Ashton, Johannes Brandstetter, and Siddhartha Mishra. Fluid Intelligence: A Forward Look on AI Foundation Models in Computational Fluid Dynamics, November 2025.

[8] Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina Agocs, Miguel Beneitez, Marsha Berger, Blakesly Burkhart, Stuart Dalziel, Drummond Fielding, et al. The Well: a large-scale collection of diverse physics simulations for machine learning. *Advances in Neural Information Processing Systems*, 37:44989–45037, 2024.

[9] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[10] Xiaoxiao Guo, Wei Li, and Francesco Iorio. Convolutional neural networks for steady flow approximation. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 481–490, 2016.

[11] Keefe Huang, Moritz Krügener, Alistair Brown, Friedrich Menhorn, Hans-Joachim Bungartz, and Dirk Hartmann. Machine learning-based optimal mesh generation in computational fluid dynamics. *arXiv preprint arXiv:2102.12923*, 2021.

[12] Neil Ashton, Charles Mockett, Marian Fuchs, Louis Fliessbach, Hendrik Hetmann, Thilo Knacke, Norbert Schonwald, Vangelis Skaperdas, Grigoris Fotiadis, Astrid Walle, Burkhard Hupertz, and Danielle Maddix. DrivAerML: High-fidelity computational fluid dynamics dataset for road-car external aerodynamics, 2025.

[13] Mohamed Elrefaie, Florin Morar, Angela Dai, and Faez Ahmed. DrivAerNet++: A large-scale multimodal car dataset with computational fluid dynamics simulations and deep learning benchmarks. *Advances in Neural Information Processing Systems*, 37:499–536, 2024.

[14] Neil Ashton, Danielle Maddix, Samuel Gundry, and Parisa Shabestari. AhmedML: High-fidelity computational fluid dynamics dataset for incompressible, low-speed bluff body aerodynamics. *arxiv.org*, 2024.

[15] Alfio Quarteroni, Paola Gervasio, and Francesco Regazzoni. Combining physics-based and data-driven models: advancing the frontiers of research with scientific machine learning. *arXiv preprint arXiv:2501.18708*, 2025.

[16] Packy McCormick and Sam D'Amico. The electric slide: The energy transition is happening faster than you think. `https://www.notboring.co/p/the-electric-slide`, May 2025. Not Boring Newsletter.

[17] Mustafa Emad, Ahmed Abdulnabi, and Sattar Aljabair. Heat transfer in electronic systems printed circuit board: A review. *Engineering and Technology Journal*, 40:99–108, 01 2022.

[18] E.R. Meinders and K. Hanjalić. Experimental study of the convective heat transfer from in-line and staggered configurations of two wall-mounted cubes. *International Journal of Heat and Mass Transfer*, 45(3):465–482, 2002.

[19] Vishwas Kulkarni and Alexander Kospach. Machine learning based reduced order model for full temperature prediction of power electronics component cooling. In *2024 Energy Conversion Congress & Expo Europe (ECCE Europe)*, pages 1–8. IEEE, 09 2024.

[20] Siemens Digital Industries Software. Simcenter STAR-CCM+, February 2025. Version 2402.0001 (19.02.013). Multiphysics CFD simulation software.

[21] Kazuto Hasegawa, Kai Fukami, Takaaki Murata, and Koji Fukagata. Machine-learning-based reduced-order modeling for unsteady flows around bluff bodies of various shapes. *Theoretical and Computational Fluid Dynamics*, 34:367–383, 2020.

[22] Nils Thuerey, Konstantin Weißenow, Lukas Prantl, and Xiangyu Hu. Deep learning methods for reynolds-averaged navier–stokes simulations of airfoil flows. *AIAA journal*, 58(1):25–36, 2020.

[23] Florent Bonnet, Jocelyn Mazari, Paola Cinnella, and Patrick Gallinari. Airfrans: High fidelity computational fluid dynamics dataset for approximating reynolds-averaged navier–stokes solutions. *Advances in Neural Information Processing Systems*, 35:23463–23478, 2022.

[24] Steeven Janny, Aurélien Beneteau, Madiha Nadri, Julie Digne, Nicolas Thome, and Christian Wolf. Eagle: Large-scale learning of turbulent fluid dynamics with mesh transformers. *arXiv preprint arXiv:2302.10803*, 2023.

[25] Neil Ashton, Jordan Angel, Aditya Ghate, Gaetan Kenway, Man Long Wong, Cetin Kiris, Astrid Walle, Danielle Maddix, and Gary Page. Windsorml: High-fidelity computational fluid dynamics dataset for automotive aerodynamics. *Advances in Neural Information Processing Systems 37*, 2024.

[26] Neil Ashton, Adam M. Clark, Christopher Ivey, Liam Heidt, Sanjeeb Bose, Rishi Ranade, Rahul Agrawal, and Konrad Goc. *High-Fidelity CFD Data Generation for HiLiftAeroML using Solution-Adapted WMLES*, page 0042. AIAA, 2026.

[27] Ante Sikirica, Luka Grbčić, and Lado Kranjčević. Machine learning based surrogate models for microchannel heat sink optimization, 2022.

[28] Vishal Anand and Mohan Sangeeth. Machine learning optimization to boost the effectiveness of phase change material (pcm)-based on-chip passive thermal management. Electronics Cooling Magazine, October 2022.

[29] Michael D McKay, Richard J Beckman, and William J Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 42(1):55–61, 2000.

[30] Consortium for Advanced Simulation of LWRs. STAR-CCM+ (CFD) Calculations and Validation L3:VVI.H2L.P15.02. Technical Report CASL-X-2017-1421-000, Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States), sep 2017.

[31] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[32] Haixu Wu, Huakun Luo, Haowen Wang, Jianmin Wang, and Mingsheng Long. Transolver: A fast transformer solver for PDEs on general geometries. *arXiv preprint arXiv:2402.02366*, 2024.

[33] Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter W Battaglia. Learning mesh-based simulation with graph networks. *arXiv preprint arXiv:2010.03409*, 2020.

[34] Adrian Wolny et al. pytorch-3dunet: 3D U-Net for volumetric segmentation, 2022.

[35] PhysicsNeMo Contributors. Nvidia physicsnemo: An open-source framework for physics-based deep learning in science and engineering. `https://github.com/NVIDIA/physicsnemo`, February 2023.

[36] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.

A Broader Impact

Contrary to many aerodynamic datasets, CoolData dataset provides a much richer diversity of geometries, making it more readily generalizable to other domains. At the same time the diversity makes it much harder to train compared to other datasets. Therefore, it constitutes a valuable resource for multiple domains, applications, and research communities in the following ways:

Despite its relatively compact geometric primitives, the dataset's structure allows it to be directly applied to selected real-world use cases, such as thermal management of PCB boards in rack-blade servers and similar computer enclosures.

- **Improving Electronics Cooling Applications:** The core focus of this dataset is the improvement of electronics cooling applications, specifically increasing their cooling efficiency. Despite its relatively compact geometric primitives, the dataset's structure allows it to be directly applied to selected real-world use cases, such as blade servers in datacenters. Since a significant portion of a datacenter energy consumption is used for cooling, CoolData will have direct impact on the energy needs of datacenters by enabling faster design iterations aiming for higher cooling efficiencies.
- **Accelerating CFD Simulations:** By providing high-fidelity simulation data, CoolData furthermore enables the training of surrogate models that substantially reduce the computational cost and time required for CFD applications well beyond electronics cooling. For example, wind simulation in cities follows similar geometries and thus the data will also impact ML models in this space. This facilitates faster design iterations in many more domains. Thereby, surrogate models may either fully replace traditional simulations or serve as an informed initial guess within an iterative solver.
- **Benchmarking and Validation:** The dataset introduces thermal management as a new benchmark domain for training and testing advanced ML models, including geometric Deep Learning (DL) approaches. It thus supports the validation and systematic comparison of different ML modeling strategies beyond external aerodynamics, which is the primary focus of benchmarking CAE ML models today.
- **Machine Learning Integration:** CoolData dataset offers a rich source for training and evaluating advanced machine learning models on complex convective heat transfer flows, supporting the broader integration of data-driven methods into engineering workflows.
- **Cross-Disciplinary Insights:** Combined with existing CFD datasets, CoolData dataset enables cross-disciplinary insights at the intersection of different flow regimes and engineering applications, fostering collaboration across research communities.

Therefore, the large dataset of 60848 electronics cooling designs is expected to drive significant innovations and advancements in both academic and industrial settings. In particular, given that the dataset reflects industrial realism through the use of an industrial-grade solver, the methodologies developed here are not necessarily limited to thermal management of electronic systems. They can be extended to thermal management applications more broadly, and even beyond, into external aerodynamics and other domains involving numerical simulations (specifically CFD) and DL.

B Dataset - Geometry Variations

The dataset consists of two to six random bodies distributed across the domain. Thereby, each body is either a cylinder or a rectangular box with position, width and height varied randomly (cf., Table 7). To investigate the variability of the data we visualize their average distribution, the corresponding variability as well as the corresponding correlations.

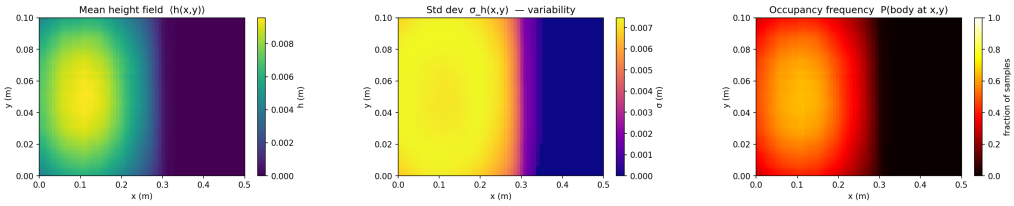
For each geometry we calculate the corresponding height field, that is the field $h(x, y) \to \mathbb{R}$ which maps to each position the height of the corresponding body. Here a height of 0

Table 7: Bounds for each parameter. Size s_x and Size s_y apply to cuboids, Radius to cylinders only. All other geometry attributes are shared between the 2 body types. Positions (x, y) refer to the centroid of cuboids and to the axis of cylinders, and the cylinder parameter denotes the radius; all bodies are mounted on the bottom wall; the cylinder parameter denotes the radius. Bodies extending beyond the lateral symmetry planes are clipped.

Parameter	Lower Bound	Upper Bound
Temperature T_{Body}	$20\,^{\circ}\text{C}$	$80\,^{\circ}\text{C}$
Inlet velocity $\vec{U}_{\text{in}}$	$1\,\text{m}\,\text{s}^{-1}$	$7\,\text{m}\,\text{s}^{-1}$
Position x	$0.05\,\text{m}$	$0.299\,\text{m}$
Position y	$0\,\text{m}$	$0.099\,\text{m}$
Radius	$0.001\,\text{m}$	$0.1\,\text{m}$
Size s_x	$0.001\,\text{m}$	$0.299\,\text{m}$
Size s_y	$0.001\,\text{m}$	$0.099\,\text{m}$
Size s_z	$0.001\,\text{m}$	$0.018\,\text{m}$

corresponds to no body being present at the specific position. Since the simulation considers a boolean combination of bodies, the height field will always indicate the maximum height, if multiple bodies are present.

In Figure 7, we visualize the mean height field $\bar{h}(x, y) = \frac{1}{N}\sum_{i=1}^{N} h_i(x, y)$, the variance $h_{\sigma^2}(x, y) = \frac{1}{N}\sum_{i=1}^{N}(\bar{h}(x, y) - h_i(x, y))^2$, and the occupancy frequency $h_{freq}(x, y) = \frac{1}{N}\sum_{i=1}^{N}[h_i(x, y) > 0]$ using the Iverson bracket notation, i.e., $[x > 0] = 1$ if $x > 0$ and 0 otherwise. As seen in Figure 7, we can observe the distribution as expected from the random sampling adopted here (cf., Section 3). That is, the mean field is higher in the middle since here geometries with center points above or below the middle contribute while closer to the boundaries there is only one sided contributions, since there are no geometries with center points outside the flow channel.



(a) Average height (b) Variance (c) Frequency

Figure 7: Distribution of heights of geometries across all samples.

To further investigate variability of the dataset, a Principal Component Analysis has been performed on set of the height functions $h_i(x, y)$, $1 \leq i \leq N$ after an appropriate discretization. The distribution of corresponding eigenvalues and the first four modes are shown in Figure 8. The slow decay of the eigenvalues as well as the major eigenmodes indicate a homogeneous distribution of geometries as expected.

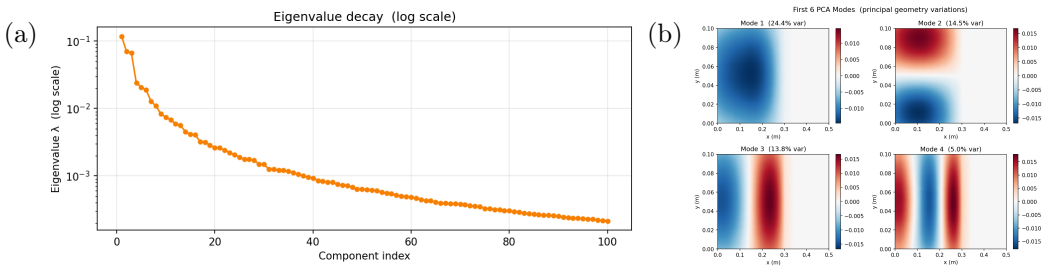


Figure 8: Principal component analysis across the $50,000$ samples of the dataset.

C Dataset - Sample Gallery

Figure 9 shows a set of 16 sample geometries from the dataset along with the visualization of different physical field variables of interest, i.e., velocity fields (streamlines), pressure fields, temperature field, and turbulent kinetic energy.

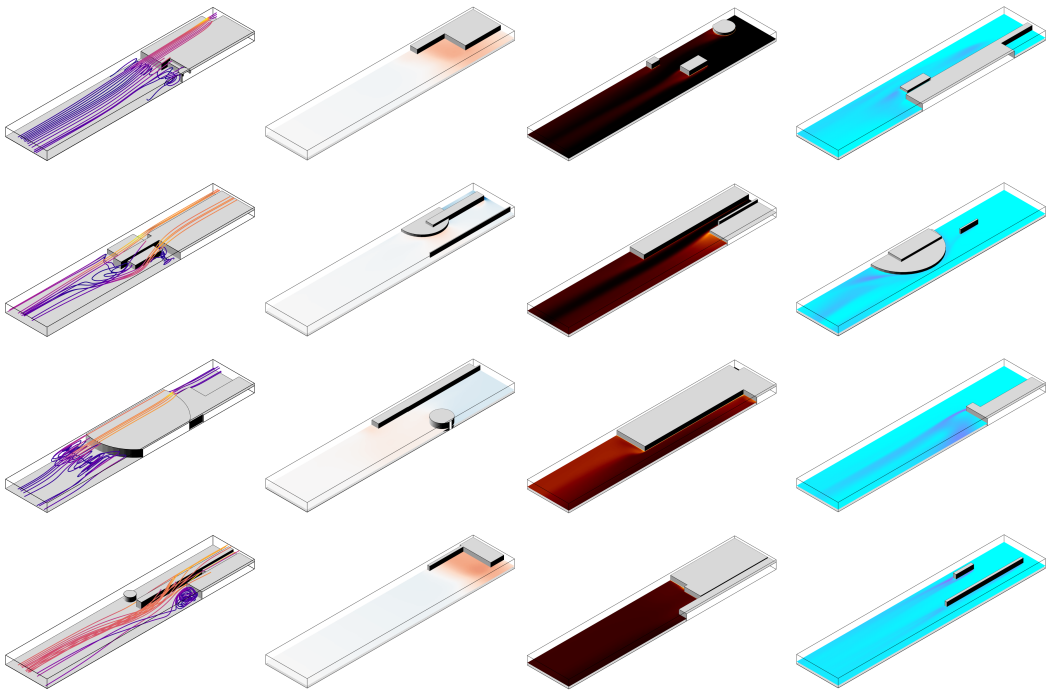


Figure 9: 16 different geometry samples from the dataset with visualization of velocity fields (streamlines), pressure fields, temperature field, and turbulent kinetic energy (columns from left to right)

D Dataset - Distribution

While the parameter space of the dataset is only 35-dimensional, its broad sampling supports out-of-distribution (OOD) studies along several independent axes. The number (2-6) and type (cuboid or cylinder) of bodies allow investigating the influence of geometric complexity (Table 8). Further axes along which the training distribution can be restricted and held-out cases evaluated are the body geometry, e.g., the body height, the thermal conditions, and the inflow velocity (Figure 11). Two of these axes are evaluated in Section 4: geometric composition, by withholding specific body-count and shape combinations from training, and thermal extrapolation, by capping the training distribution's peak body temperature. The remaining axes are not evaluated in this paper, but their coverage shown in Figure 11 demonstrates that CoolData supports OOD study designs along each of them.

Table 8: Number of simulations by active body count, the parameter underlying this paper's geometric-composition OOD study (Section 4).

Active Bodies	With Cylinder	Cuboids Only	Total
2	0	26,525	26,525
3	12,184	12,076	24,260
4	7,184	1,441	8,625
5	1,342	0	1,342
6	96	0	96

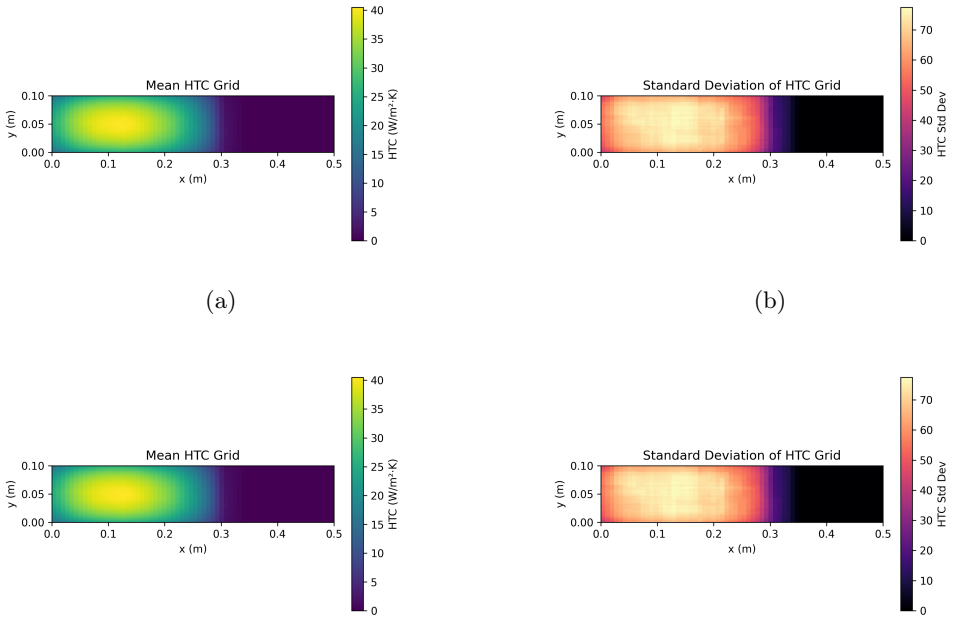
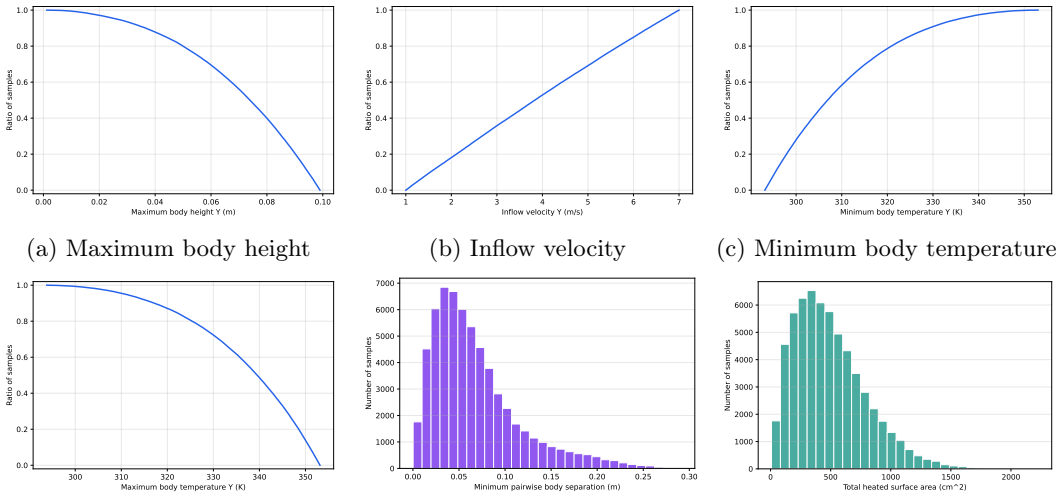


Figure 10: Cumulative frequencies of continuous simulation parameters varied across the dataset: (a) maximum body height, (b) minimum body temperature, (c) maximum body temperature, and (d) inflow velocity.



(a) Maximum body height (b) Inflow velocity (c) Minimum body temperature

(d) Maximum body temperature (e) Minimum body separation (f) Total heated surface area

Figure 11: Distributions of candidate OOD axes across the dataset: (a)-(d) cumulative distributions of continuous simulation parameters, (e)-(f) histograms of derived geometric quantities. Panel (d) underlies the thermal-extrapolation study in Section 4; all other panels show axes not evaluated in this paper.

E Dataset - In-Depth Insights

Figure 12 shows a comparative analysis of selected hydraulic performance metrics such as the heat dissipated, mean HTC coefficient (across all surfaces), inlet velocity, and maximum temperature.

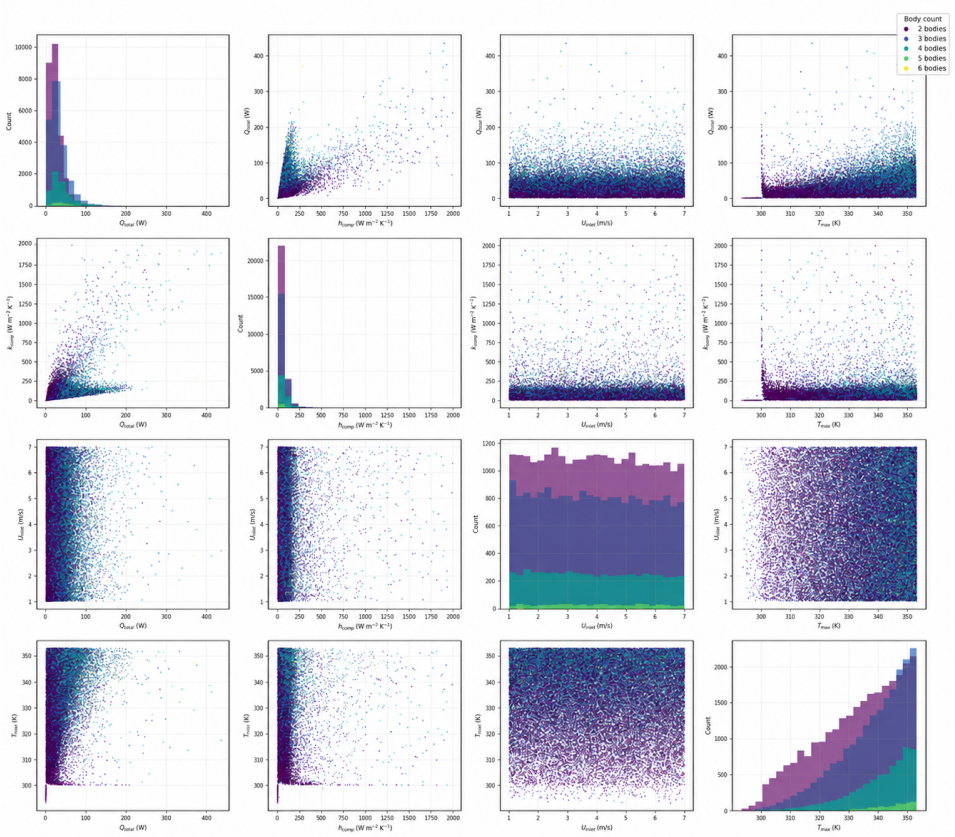


Figure 12: Pair plot of key thermal performance metrics in the dataset. The diagonal histograms show the marginal distributions of total heat dissipated Q_{total}, mean component heat transfer coefficient $\bar{h}_{\text{comp}}$, inlet velocity U_{inlet}, and maximum component temperature T_{max}. The off-diagonal scatter plots visualize pairwise relationships between these quantities, revealing the trade-offs between cooling performance, thermal safety, and flow conditions. Colors denote geometry categories based on body count, illustrating how geometric complexity influences the occupancy of the performance space and highlighting the diversity of configurations represented in the dataset.

In Figure 13, the scatter plot reveals that the majority of samples cluster at low mean HTC values (below 500 W/m^{2}K), consistent with realistic forced convection conditions in electronics cooling applications, while maximum HTC spans a much wider range up to 10,000 W/m^{2}K, indicating highly localized heat transfer hotspots. The colour encoding shows a clear positive correlation between body temperature and both mean and maximum HTC, confirming that higher thermal loads drive more intense convective transfer. This distribution demonstrates the dataset's physical consistency and the challenge it poses for surrogate models: capturing rare but physically significant high-HTC outliers alongside the dense low-HTC majority.

In Figure 14 we consider the spatial distribution of the heat transfer coefficient. This nicely correlates with the geometry distribution as shown in Figure 7 as expected. Figure 15 shows the distribution of key surface quantities across the dataset. The Max HTC and Mean HTC distributions are right-skewed, with most samples concentrated at lower values and a long tail of high-intensity outliers, consistent with the scatter plot in Figure 13. The Total Heat Dissipated distribution similarly peaks at low values. The body temperature distributions

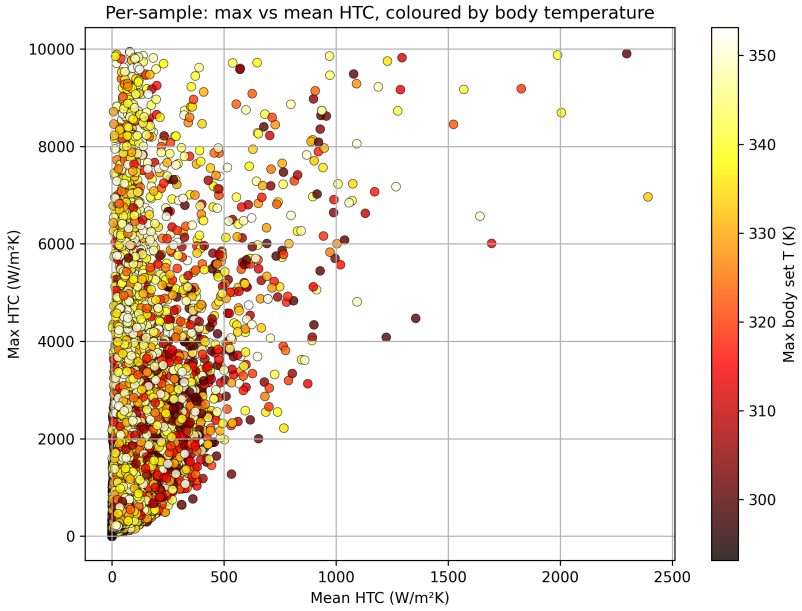
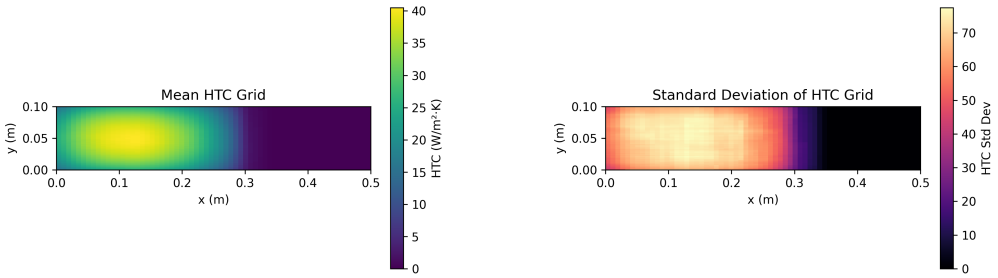


Figure 13: Per-sample relationship between mean and maximum wall HTC, coloured by the maximum body set temperature



(a) (b)

Figure 14: Spatial distribution of heat transfer coefficients and their variation. (a) Mean heat transfer coefficient and (b) standard deviation.

(mean, max, min) span approximately 290-355 K with a relatively uniform spread, indicating good coverage of the operating temperature range and avoiding bias toward any particular thermal regime.

F Deep Surrogate Models

The provided dataset is benchmarked on 4 different DL models, namely 3D U-Net (31), MeshGraphNet (33), Transolver (32), and DOMINO (6), along two complementary axes: a volumetric benchmark (3D U-Net, Transolver, DOMINO) predicting the temperature field, and a surface-centric benchmark (MeshGraphNet, Transolver, DOMINO) predicting the wall shear stress magnitude. To do so, we rely on openly available implementations of the corresponding models: NVIDIA's `PhysicsNeMo` (35) for MeshGraphNet, Transolver, and DOMINO, and `pytorch-3dunet` (34) for 3D U-Net.

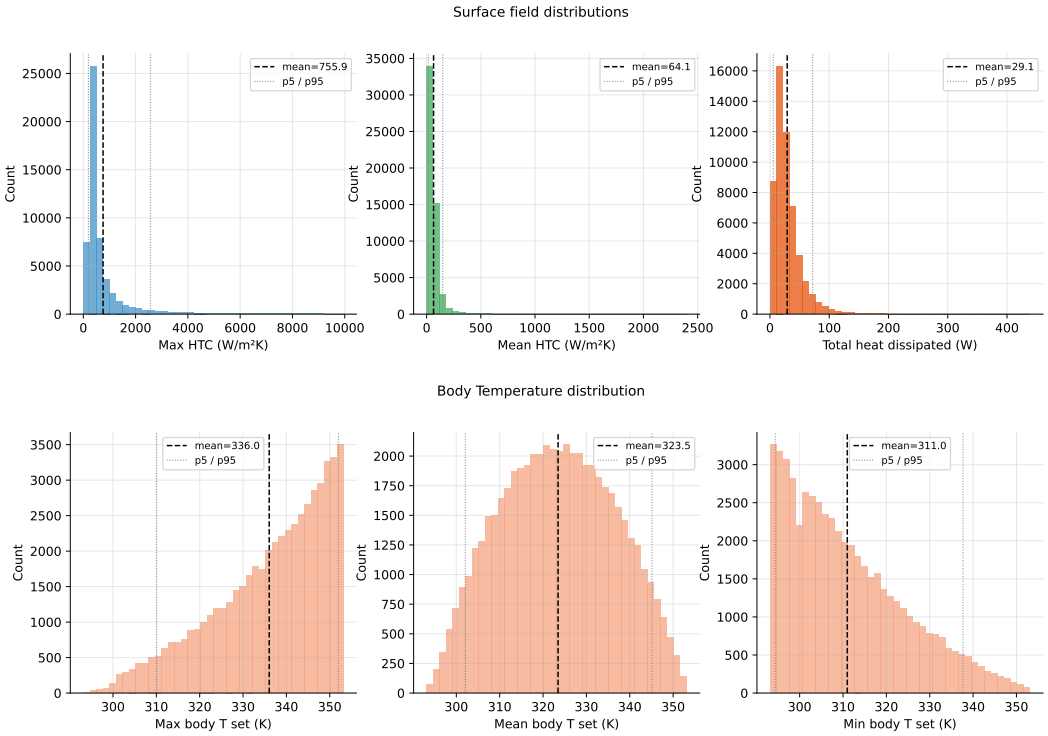


Figure 15: Distribution of key surface quantities (Total Heat Dissipated, Mean and Max HTC, as well as Mean, Max, Min body temperature) across samples.

3D U-Net

We evaluate a 3D U-Net (31) with FiLM (36) meta-conditioning as a grid-based baseline for temperature field prediction. Unlike point-based neural operators, the model operates on a structured volumetric representation of the physical domain.

Problem formulation: The input domain is discretized into a regular grid of resolution 64^3 and the model learns a mapping

$$\mathbf{X} \in \mathbb{R}^{4\times D\times H\times W} \rightarrow \mathbf{Y} \in \mathbb{R}^{1\times D\times H\times W},$$

where $D = H = W = 64$. The domain ($0.5 \times 0.1 \times 0.02$ m) is non-cubic, so this fixed voxel count per axis corresponds to anisotropic voxel spacing ($\Delta x \approx 7.9$ mm, $\Delta y \approx 1.6$ mm, $\Delta z \approx 0.3$ mm), rather than isotropic cubic voxels. Being based on a volumetric discretization, 3D U-Net is effectively a volume model.

Input representation: The input tensor is constructed from four channels, occupancy, body temperature, inflow condition, and an approximate signed distance field (SDF) together with a separate 35-dimensional global parameter vector injected via FiLM conditioning (see below) rather than voxelized onto the grid directly. Occupancy and body temperature are rasterized directly from the analytic geometric primitives (axis-aligned bounding-box tests for cuboids, radial/axial tests for cylinders); the inflow channel is set to a spatially constant value on the single grid slice at the domain's inflow boundary, rather than broadcast across the full volume.

Geometry and physics encoding: The SDF channel is computed as a discrete approximation, not an exact analytic distance field: a Euclidean distance transform is applied separately to the occupied and unoccupied regions of the voxelized occupancy grid, and the two are combined (positive inside solid bodies, negative in the fluid domain) and normalized to $[-1, 1]$. This is a resolution-limited approximation and, together with the grid-based

rasterization of the boundary conditions above, means the model's effective input geometry is only as precise as the 64^3 voxelization itself, not the original analytic CAD-level geometry. Similarly, the ground-truth temperature label is prepared by projecting the original, unstructured CFD cell centers onto the regular grid via nearest-neighbor lookup (a k-d tree query per voxel); this introduces a discretization error into the target itself, independent of any model error, meaning the reported metrics for 3D U-Net reflect deviation from the *voxelized* ground truth rather than the original CFD field directly.

Global simulation parameters (inflow velocity, and up to 6 heat source objects' position, temperature, and size) are encoded into a 35-dimensional vector identical in structure to the global parameter encoding used by Transolver's and DOMINO's volumetric variants and injected into the network via Feature-wise Linear Modulation (FiLM): a small MLP maps this vector to per-channel scale and shift parameters (γ, β), which modulate the U-Net's feature maps as $\mathbf{F}' = \mathbf{F} \cdot (1 + \gamma) + \beta$ after the convolutional encoder-decoder, allowing the network to condition its volumetric prediction on global, non-spatial simulation parameters that are not naturally expressed as a voxel channel.

Model architecture: We use a 3D U-Net with 3 encoding levels and a base feature width of 35, with FiLM conditioning applied after the convolutional backbone via a 2-layer MLP (hidden size 64). The network takes a 4-channel volumetric input and a 35-dimensional global parameter vector, and produces a single-channel temperature field output. Standard convolutional downsampling with skip connections is used to preserve spatial detail, without a final sigmoid activation, as the target is a continuous temperature field rather than a segmentation mask. The model contains 1,184,978 trainable parameters.

MeshGraphNet

We evaluate MeshGraphNet (33) as a graph-based neural operator for predicting wall shear stress on unstructured surface meshes.

Problem formulation: Each sample is represented as a graph $G = (V, E)$ constructed from body-surface mesh nodes, restricted to the heated-body ("Pipe Body") surfaces, not the full channel/domain boundary, the same restriction used by Transolver's surface-centric variant. The model learns a mapping from node- and edge-level features to a scalar wall shear stress field:

$$(\mathbf{X}_V, \mathbf{X}_E, G) \to \mathbf{Y}_V,$$

where $\mathbf{X}_V \in \mathbb{R}^{N \times 41}$ are node features, $\mathbf{X}_E \in \mathbb{R}^{E \times 3}$ are edge features, and $\mathbf{Y}_V \in \mathbb{R}^{N \times 1}$ is the target field. Taking only body-surface meshes as input, MeshGraphNet is effectively a surface-based model, matching Transolver's and DOMINO's surface-centric variants on the same benchmark.

Node representation: Each node's 41-dimensional feature vector is the concatenation of its (normalized) spatial position (3), outward surface normal (3), and a 35-dimensional global parameter vector (inflow velocity, and up to 6 heat source objects' temperature, position, and size) broadcast to every node, the identical per-point encoding used by Transolver's surface-centric variant (see Transolver subsection above), enabling a fair architecture comparison on the same input representation.

Edge representation: Edges are defined by mesh connectivity, and each edge is associated with a normalized displacement vector:

$$\mathbf{x}_{ij} = \mathbf{p}_j - \mathbf{p}_i,$$

scaled by the domain diagonal to ensure numerical stability. This encodes local geometric interactions between neighbouring surface elements.

Model architecture: MeshGraphNet consists of a node encoder, edge encoder, processor, and decoder, each implemented as 2-layer MLPs with hidden dimension 128. The processor performs 8 message passing steps with sum aggregation. Node and edge inputs are 41- and

3-dimensional respectively, and the output is a scalar per node. The model contains 1,295,489 trainable parameters.

Transolver

We evaluate Transolver (32) as a transformer-based neural operator for learning mappings between functional inputs and solution fields on unstructured point clouds. Transolver is evaluated on both the surface-centric (wall shear stress) and volumetric (temperature) benchmarks, using two separate input representations described below.

Problem formulation: Each sample consists of an unstructured point cloud representing a 3D surface with N points. The model learns a mapping

$$(\mathbf{p}, \mathbf{u}) \to \mathbf{y},$$

where $\mathbf{p} \in \mathbb{R}^{N\times 3}$ denotes spatial coordinates, $\mathbf{u} \in \mathbb{R}^{N\times 40}$ denotes point-wise input features, and $\mathbf{y} \in \mathbb{R}^{N\times 1}$ is the target temperature field. In the context, of our contribution we focus on surface points since these are the main points of interest. Thus, Transolver is effectively a surface model.

Surface-centric variant (wall shear stress): Each sample consists of an unstructured point cloud of N body-surface query points, restricted to the heated-body ("Pipe Body") surfaces, not the full channel/domain boundary. The model learns a mapping

$$(\mathbf{p}, \mathbf{u}) \to \mathbf{y},$$

where $\mathbf{p} \in \mathbb{R}^{N\times 3}$ denotes spatial coordinates, $\mathbf{u} \in \mathbb{R}^{N\times 41}$ denotes point-wise input features, and $\mathbf{y} \in \mathbb{R}^{N\times 1}$ is the target wall shear stress field. Each point's 41-dimensional input feature is the concatenation of its (normalized) spatial position (3), outward surface normal (3), and the same 35-dimensional global parameter vector broadcast to every point, following the identical encoding used throughout the other CoolData pipelines. The target wall shear stress magnitude is normalized via $y = \log(1 + \mathrm{WSS}_{\mathrm{clip}})/\log(1 + \mathrm{WSS}_{\mathrm{max}})$, with $\mathrm{WSS}_{\mathrm{clip}} = \min(\mathrm{WSS}, \mathrm{WSS}_{\mathrm{max}})$ and $\mathrm{WSS}_{\mathrm{max}} = 15.0$ Pa, compressing the field's naturally heavy-tailed dynamic range before training; predictions are inverse-transformed back to Pa for evaluation.

Volumetric variant (temperature): Each sample consists of an unstructured point cloud of N volume cell centers, sampled from the CFD mesh's internal fluid domain. The model learns the same functional mapping $(\mathbf{p}, \mathbf{u}) \to \mathbf{y}$, but here $\mathbf{u} \in \mathbb{R}^{N\times 35}$ (no surface normals, since interior points have none) and $\mathbf{y} \in \mathbb{R}^{N\times 1}$ is the target temperature field. The 35-dimensional input encodes the normalized inflow velocity together with up to 6 heat source objects' temperature, position, and size, following the same encoding scheme as the surface variant but without the normal-vector and surface-type components. During training, each volume sample is subsampled to a random 10,000 points per epoch, drawn fresh each epoch (see Training, below).

Model architecture: Both variants use the same underlying architecture: 6 Transolver blocks with hidden dimension 128 and 8 attention heads, a physics token space discretized into 128 slices, and spatial coordinates embedded into a 3-dimensional positional encoding. The surface-centric model contains 1,153,073 trainable parameters; the volumetric variant contains 1,151,537 trainable parameters, the difference owing entirely to the differing input dimensionality (41 vs. 35).

DOMINO

We evaluate DOMINO (6), a decomposable multi-scale neural operator, on both the surface-centric (wall shear stress) and volumetric (temperature) benchmarks. Like Transolver, DOMINO operates on unstructured point clouds rather than a fixed grid; unlike Transolver, it combines this point-based representation with an explicit multi-scale geometry encoding derived from the surface STL mesh, applied identically in structure to both variants but with different sampling and radii, described below.

Problem formulation: DOMINO learns a mapping from query points $\mathbf{p} \in \mathbb{R}^{N\times 3}$, together with a 35-dimensional global parameter vector (the same global simulation-parameter encoding used throughout CoolData's pipelines), to a scalar target field $\mathbf{y} \in \mathbb{R}^{N\times 1}$. The two variants differ in which points constitute $\mathbf{p}$ and $\mathbf{y}$, and in their per-epoch subsampling:

Volumetric variant (temperature): $\mathbf{p}$ are volume cell centers and $\mathbf{y}$ is the temperature field. During training, N is subsampled to 10,000 volume points per epoch, matching Transolver's volumetric subsampling strategy; a separate 10,000 points are sampled from the surface STL geometry for the geometry encoder.

Surface variant (wall shear stress): $\mathbf{p}$ are body-surface query points, restricted to the heated-body ("Pipe Body") surfaces as in Transolver's and MeshGraphNet's surface variants, and $\mathbf{y}$ is the wall shear stress field. During training, N is subsampled to 999 surface points per epoch; a separate 10,000 points are sampled from the surface STL geometry for the geometry encoder.

Geometry encoding: Geometric information is encoded via *two* complementary mechanisms operating at multiple spatial scales, applied to both variants:

*Multi-scale geometry representation (*geo_conv + geo_processor*):* A geometry convolution aggregates local neighborhood information around each point at three radii, considering up to 10 neighbors within each radius. These radii differ between variants: $(0.02, 0.1, 0.5)$ for the volumetric variant, $(0.005, 0.02, 0.05)$ for the surface-centric variant (normalized domain units). This is followed by a convolutional geometry processor (8 base filters) that refines the multi-scale representation; DOMINO additionally incorporates a signed distance field directly into its basis function computation (rather than as a separate input channel, as in 3D U-Net).

*Local point-neighborhood encoding (*geometry_local*):* A second, complementary encoder aggregates local geometric information at two additional radii per point, considering up to 32 and 128 neighbors respectively. These radii also differ between variants: $(0.02, 0.1)$ for the volumetric variant, $(0.01, 0.05)$ for the surface-centric variant.

Each query point's prediction is computed using its 7 nearest neighbors, for both the volumetric and surface-centric variants.

Model architecture: The basis function network uses a base hidden layer of 128 with Fourier feature positional encoding (5 modes). Separate 128-dimensional networks encode spatial position and the global parameter vector, and a final 128-dimensional aggregation network combines the geometry, basis-function, and parameter encodings to produce the point-wise prediction. The volumetric and surface-centric variants are trained as separate model instances (not jointly). The model contains 1,196,807 trainable parameters for the volumetric variant and 1,144,475 for the surface variant.

Training

All models are trained on 8 NVIDIA H100 GPUs using distributed data parallel training with the AdamW optimizer, for 100 epochs. 3D U-Net, MeshGraphNet, and Transolver use a fixed learning rate of 1×10^{-4} with ReduceLROnPlateau scheduling (reduction factor 0.5, patience 5 epochs, triggered on validation-loss plateaus). DOMINO uses a different learning rate and schedule: an initial learning rate of 1×10^{-4} with a fixed-milestone step decay (MultiStepLR, decay factor 0.5) rather than an adaptive, validation-triggered schedule.

Mean squared error (MSE) between predicted and ground-truth field values serves as the training objective for all four models. Mixed-precision training and gradient clipping are applied for numerical stability and efficiency.

To assess the effect of dataset size on model performance, each model is trained at five scales: $1,000$, $5,000$, $10,000$, $25,000$, and $50,000$ samples.

A fixed set of $1,000$ validation and $1,000$ test samples is reserved before any training subset is drawn, and held constant across all scales; the remaining pool of samples is available

for training. The scaling is cumulative: the training samples used at a smaller scale are a strict subset of those used at every larger scale, ensuring that performance differences reflect the addition of new training data rather than changes in the training distribution or validation/test composition. Evaluation is performed on the fixed held-out test set of $1,000$ samples, not seen during training at any scale.

MeshGraphNet operates on variable-length surface meshes, precluding standard batching; the DataLoader prefetches 48 samples to keep the GPU pipeline saturated, while the model processes one sample per forward pass. Transolver's volumetric variant subsamples every sample to a fixed 10,000 points per epoch, drawn fresh each epoch, enabling true batching (batch size 8 per GPU, via a wrapper that stacks same-shaped subsampled point clouds); the surface-centric variant trains on the full, unmodified point cloud with batch size 1 and no worker-process prefetching. DOMINO's volumetric variant also subsamples to 10,000 volume points per epoch (plus a separate 10,000 points from the surface geometry for its geometry encoder), matching Transolver's volumetric subsampling target, but is trained with batch size 1 rather than exploiting the fixed-size representation for larger-batch training (a limitation of the DOMINO model).

MeshGraphNet, Transolver, and DOMINO are implemented via NVIDIA PhysicsNeMo (35), and 3D U-Net via `pytorch-3dunet` (34). Specific hyperparameters used are shown in Table 9 and an exemplary plot of the loss function is shown in Figure 16.

Table 9: Training hyperparameters for all models. All models are trained with distributed data parallel (DDP) on 8 NVIDIA H100 GPUs with AdamW optimisation for 100 epochs.

Hyperparameter	**3D U-Net**	**MeshGraphNet**	**Transolver**	**DOMINO**
Epochs	100	100	100	100
Batch size	16	1 (prefetch 48)	1 surface / 8 volumetric	1 (8-step gradient accumulation)
Learning rate	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}
LR scheduler	ReduceLROnPlateau (factor=0.5, patience=5)			MultiStepLR (milestones=[100,200,350], $\gamma = 0.5$)
Gradient clip	1.0	1.0	1.0	1.0
Loss function	MSE	MSE	MSE	MSE
Mixed precision	FP16	FP16	FP16	FP16
Weight decay	10^{-2}	10^{-2}	10^{-2}	10^{-2}
Model selection	Best val. MSE	Best val. MSE	Best val. MSE	Best val. loss

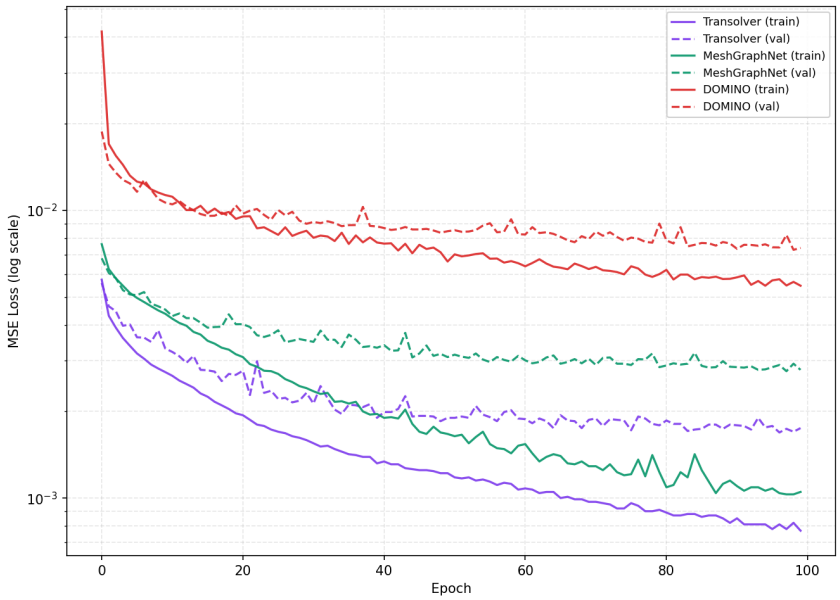


Figure 16: Visualization of loss function for all three surface models trained on a dataset with 50,000 samples and validated on 1000 held-out samples. Absolute loss values are not directly comparable across models due to model-specific normalization and point sampling.

G Evaluation

While all models in this study are trained using either a Mean Squared Error (see Table 9), we use additional quantities to evaluate model performance. The key metrics used in this study are:

Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(T_i - \hat{T}_i)^2$$

Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|T_i - \hat{T}_i|$$

Maximum Absolute Error (MaxAE):

$$\text{MaxAE} = \max_{i\in\{1,n\}}|T_i - \hat{T}_i|$$

Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(T_i - \hat{T}_i)^2}{\sum_{i=1}^{n}(T_i - \bar{T}_i)^2}$$

with $\bar{T}_i = \frac{1}{n}\sum_{i=1}^{n} T_i$. Each metric above is computed as a single pooled value over every point in the test set (not averaged per-sample), so that samples with more points contribute proportionally more to the reported value.

Models are compared within two separate groups, matching the two benchmarks described above: 3D U-Net, Transolver, and DOMINO are compared against each other on the volumetric (temperature) benchmark, and MeshGraphNet, Transolver, and DOMINO are compared against each other on the surface-centric (wall shear stress) benchmark. Correspondingly, each model's evaluation metrics are computed over the representation native to its own benchmark. For 3D U-Net, error is evaluated voxel-wise, i.e., the true result is likewise transformed into a voxel-based representation before comparison. For MeshGraphNet, error is evaluated across all nodes of the surface discretization. For Transolver and DOMINO, error is evaluated over all points of the corresponding point cloud in each benchmark they participate in the surface discretization for the wall-shear-stress benchmark, and the full-resolution volume cell centers (not subsampled, unlike during training) for the temperature benchmark.

H Body-Composition OOD Results for Further Models

Figures 19 and 20 show the body-composition OOD learning curves on the volumetric benchmark for UNet3D+FiLM and DOMINO, complementing the Transolver results in Figure 4. The setup is identical: all models are trained exclusively on the 3q0c composition, and all other compositions are held out entirely from training. Exact values at $N_{\text{train}} = 11{,}000$ for all three models are reported in Table 3. The body-composition study on the surface benchmark (Figure 6) was conducted for Transolver only.

I Notation and Main Physical Quantities

The dataset provides a comprehensive set of physical quantities commonly used in CFD. These are divided into volume and surface fields, each represented on an unstructured mesh and stored in the `CGNS` format. Units are provided in SI.

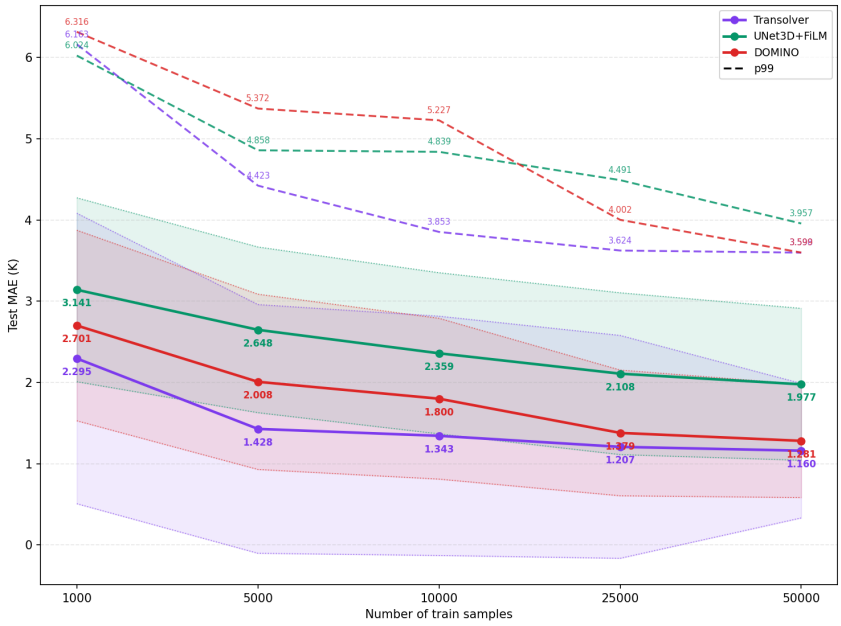


Figure 17: Visualization of temperature learning curve for all three volume models.

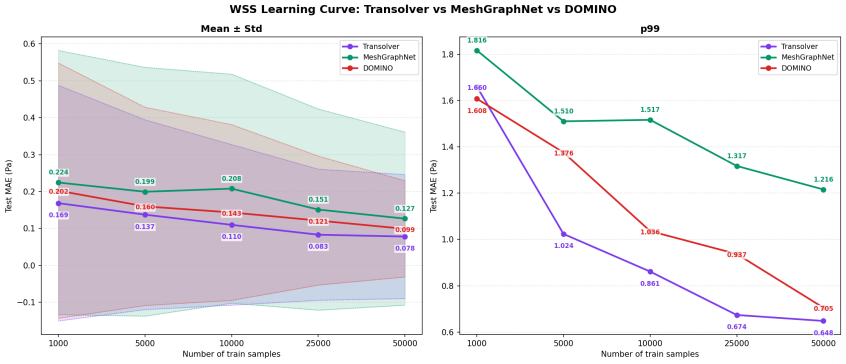


Figure 18: Visualization of WallShearStressMagnitude learning curve for all three surface models .

Table 10: Relative error (MAE as a percentage of the mean signal magnitude, 5.2 K) for the temperature-based OOD studies, complementing the wall-shear-stress relative-error result in the main text (Table 6). Unlike WSS, in-distribution relative error is modest for both studies here; degradation is driven primarily by extrapolation distance rather than intrinsic task difficulty.

	Body-composition (temperature), $N = 11{,}000$				
Metric	**3q0c**	**2q1c**	**2q0c**	**4q0c**	**4q2c**
MAE (K)	1.20	2.68	0.99	3.06	4.70
Relative error (%)	23.1	51.5	19.0	58.8	90.4

	Thermal extrapolation, $N = 15{,}000$		
Metric	**A (in-dist.)**	**B (moderate)**	**C (severe)**
MAE (K)	0.84	1.54	2.26
Relative error (%)	16.2	29.6	43.5

Volume Field Quantities

Each volume sample contains the following fields:

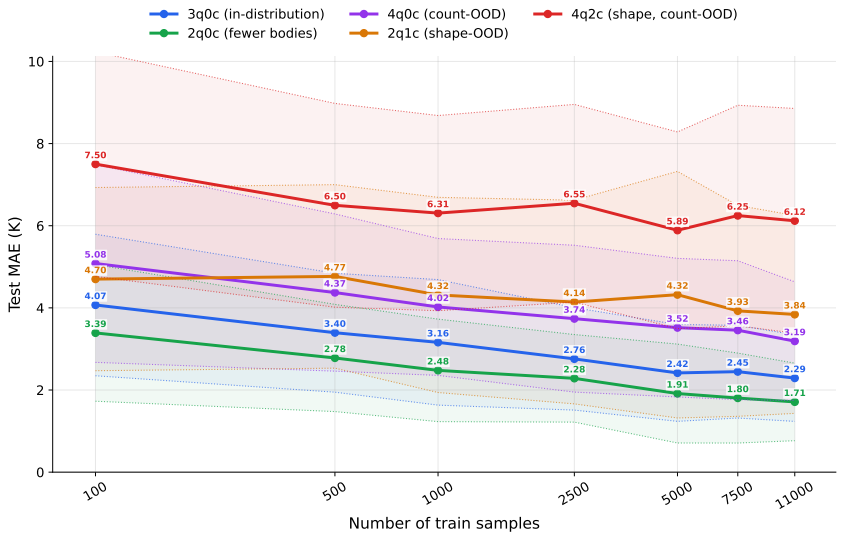


Figure 19: Body-composition OOD learning curves for UNet3D+FiLM on the volumetric benchmark: temperature test MAE (mean ± std, shaded) versus number of training samples, for five body compositions. 3q0c is in-distribution; all other compositions are held out entirely from training. For reference, the mean temperature rise above the inlet across the dataset is 5.2 K.

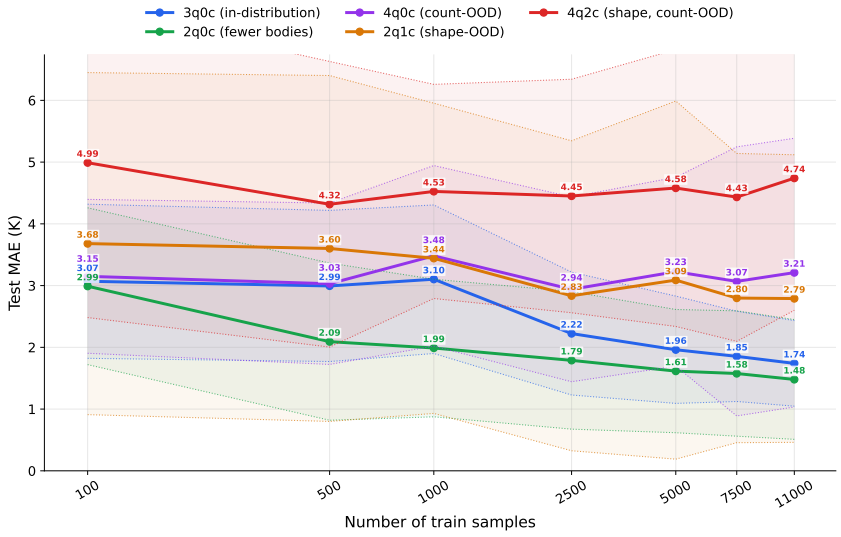


Figure 20: Body-composition OOD learning curves for DOMINO on the volumetric benchmark: temperature test MAE (mean ± std, shaded) versus number of training samples, for five body compositions. 3q0c is in-distribution; all other compositions are held out entirely from training. For reference, the mean temperature rise above the inlet across the dataset is 5.2 K.

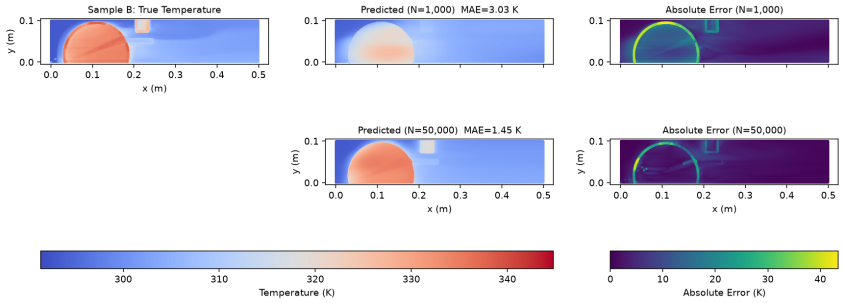


Figure 21: Comparison of true (CFD ground truth) and Transolver-predicted volumetric temperature fields for a representative test sample, evaluated at two training set sizes (1000 and 50000).

- `Velocity` $\vec{v}$ (m/s): A 3D vector field representing the fluid velocity components in the x, y, and z directions (`Velocity_0`, `Velocity_1`, `Velocity_2`).
- `Pressure` p (Pa): The static pressure field.
- `Temperature` T (K): The temperature of the fluid.
- `TurbulentKineticEnergy` k ($\mathrm{m}^2/\mathrm{s}^2$): The specific turbulent kinetic energy.
- `TurbulentDissipationRate` ε ($\mathrm{m}^2/\mathrm{s}^3$): The dissipation rate of turbulent kinetic energy.

Surface Field Quantities

Each surface sample includes wall-related quantities defined on the geometry surface:

- `AreaMagnitude` A (m^2): The surface area associated with each wall element.
- `HeatTransferCoefficient` h ($\mathrm{W}/(\mathrm{m}^2 \cdot \mathrm{K})$): The local convective heat transfer coefficient at the wall.
- `Normal` $\vec{n}$ (unitless): A unit 3D vector normal to the surface, with components `Normal_0`, `Normal_1`, and `Normal_2`.
- `Temperature` T_{body} (K): The wall temperature, which is constant across a body.
- `Pressure` p (Pa): The static pressure on the wall surface.
- `WallShearStressMagnitude` $|\vec{\tau}_{\mathrm{w}}|$ (Pa): The magnitude of the wall shear stress vector.
- `WallShearStress` $\vec{\tau}_{\mathrm{w}}$ (Pa): A 3D vector representing wall shear stress components in the x, y, and z directions (`WallShearStress_0`, `WallShearStress_1`, `WallShearStress_2`).

Derived Heat Transfer Related Quantities

Based on the saved volumetric and surface data, appropriate engineering related quantities can be derived, e.g., taking appropriate surface integrals. Below we summarize the heat transfer related quantities we use for evaluation of the dataset:

- Mean HTC $\bar{h}_{\mathrm{aw}}$ ($\mathrm{W\,m^{-2}\,K^{-1}}$): Area-weighted mean surface-integral, i.e., $\bar{h}_{\mathrm{aw}} = \int h \,\mathrm{d}A / \int \mathrm{d}A$.
- Body-to-inlet ΔT_{body} (K): Temperature difference between inflow and body, i.e., $\Delta T_{\mathrm{body}} = T_{\mathrm{body}} - T_{\mathrm{in}}$.
- Body heat dissipation Q_{body} (W): Heat flux integrated over a single body, i.e. $Q_{\mathrm{body}} = \sum_{i \in \mathrm{faces}} h_i A_i \Delta T_{\mathrm{body}}$.
- Total heat dissipation Q_{total} (W): Sum of Q_{body} over all heated bodies in a sample.