

Dimensionally Constrained Symbolic Regression over Thermal-Hydraulic Dimensionless Groups

Byung-Hyun You *

Korea Atomic Energy Research Institute, 111, Daedeok-daero 989beon-gil, Yuseong-gu, Daejeon 34057,
Republic of Korea

*Corresponding author: bhyou@kaeri.re.kr

***Keywords:** symbolic regression, thermal hydraulics, heat transfer coefficient, empirical correlation

1. Introduction

In two-phase flow and boiling systems, engineering quantities are commonly predicted using empirical correlations formulated in terms of dimensionless parameters. Symbolic regression is a machine-learning method that discovers human-interpretable symbolic models from data [1]. It eliminates the need to assume a functional form in advance; however, an expression assembled freely from operators is not guaranteed to be dimensionally consistent with the target variable. This paper presents a methodology that enforces dimensional consistency by construction: the regression target is made dimensionless, every candidate feature is a thermal-hydraulic dimensionless group, and any expression built from these features is therefore dimensionally consistent with the target. The symbolic-regression library itself is used without modification: the dimensional constraint enters entirely through the dimensionless target and the candidate feature set, while a separate constraint on the power operator restricts its exponent to a numeric constant.

The dimensional constraint is general, whereas the field-specific physics is encoded in the candidate feature set. For the thermal-hydraulic problems considered here, this set consists of dimensionless parameters established in the two-phase flow and boiling literature, ratios and normalized differences between variables with the same dimensions, and products of integer powers of the base variables whose dimensions cancel. The methodology is demonstrated for saturated flow boiling heat transfer in helically coiled tubes, and the selected candidate correlations are reported together with their accuracy and complexity.

2. Methodology

2.1 Dimensional constraints

All base variables are expressed in SI base units, and all fluid properties are evaluated at saturation, so that each property is a function of pressure alone. The regression target is the ratio h/h_l of the two-phase heat transfer coefficient to the single-phase heat transfer coefficient with only the liquid fraction flowing in the tube. The target and all candidate features are

dimensionless; any expression constructed from them is therefore dimensionally consistent with the target.

2.2 Candidate feature set

The candidate feature set consists of 45 dimensionless groups in five categories: (i) ratios of variables with the same dimensions; (ii) normalized differences; (iii) dimensionless flow-state variables; (iv) dimensionless parameters established in the two-phase flow and boiling literature, including liquid- and vapor-phase forms, the boiling Reynolds numbers, and the Lockhart-Martinelli parameter, included as $1/X_{tt}$ because the two-phase convective enhancement increases as X_{tt} decreases; and (v) products of the base variables raised to bounded integer powers whose combined dimensions cancel. A product is excluded if it can be composed from simpler retained products or is strongly correlated with the retained groups. Here, d is the tube inner diameter, D_c the coil diameter, and p_c the coil pitch. For straight-tube data, the geometric ratios vanish and the normalized geometric differences reduce to unity, so that the same feature set covers the zero-curvature limit. The candidate feature set is constructed without reference to the measured heat transfer coefficients.

2.3 Symbolic-regression search

The search uses SymbolicRegression.jl [1] with the operators restricted to $+$, $-$, $\times$, $\wedge$, and $\sqrt{}$. For the power operator, an operator-specific constraint leaves the base unrestricted but limits the exponent to a single node. Any candidate whose exponent is not a numeric constant is assigned an infinite loss and rejected during the search. Every discovered exponent is therefore a real constant, optimized together with the other scalar constants in the evolve-simplify-optimize loop. A total of 64 independent runs are performed. Pareto-optimal expressions are collected along the accuracy-complexity front, where the complexity of an expression is the number of nodes in its expression tree.

2.4 Factored forms

Each discovered expression is expanded into a sum of products of the candidate features with constant

coefficients and exponents. When the numerical constants are removed, expressions that differ only by algebraic rearrangement collapse onto a common factored form. Each term is then characterized by its effective exponents of heat flux and mass flux. A candidate is retained if its prediction increases monotonically with both the boiling number and $1/X_{tt}$. It must also include a convective contribution independent of the heat flux, so that the prediction does not vanish as $q'' \rightarrow 0$.

3. Demonstration

3.1 Database and objective

A database of 2,839 data points was digitized in this work from the 12 sources compiled[2]: seven sources with helically coiled tubes (634 data points) and five with straight tubes. The vapor quality of the data ranges from 0.005 to 0.95. Of these, 22 data points were excluded because their measured heat transfer coefficients differed by more than a factor of 2.5 from the median prediction of four classical correlations [3-6]. The remaining data were then limited to $P \leq 8$ MPa, $G \leq 600$ kg/m²s, and $q'' \leq 300$ kW/m², leaving a filtered database of 941 data points. The loss function is the squared relative error in h/h_i , with the weights chosen so that each cell of a $P \times (h/h_i)$ partition of the filtered database contributes equally to the objective. This weighting emphasizes the 2-8 MPa pressure range relevant to the steam generators of small modular reactors.

3.2 Results

The 64 runs produced 852 distinct Pareto-optimal expressions whose RMS relative error over the filtered database does not exceed 0.2; every run yielded at least one such expression. After the constants were removed, the 852 expressions collapsed onto 572 factored forms. The criteria of Sec. 2.4 then reduced these to a small set of candidates: a family in which a power of the sum of the boiling number and $1/X_{tt}$ multiplies the coil-pitch group; a superposition form in which convective and heat-flux contributions are added; and a two-branch form. The selected candidates are listed in Table I. The most compact expression of these is

$$h/h_i = \frac{p_c - d}{p_c} \left(\frac{1}{X_{tt}} + 244.93\sqrt{Bo} \right) \quad (1)$$

An expression with the same structure and a nearly identical constant was obtained in 15 of the 64 runs. Figure 1 compares the predictions of Eq. (1) with the measured data. Over the filtered database, the RMS relative error is 0.170 and 93.5% of the data are predicted within $\pm 30\%$.

Table I. Selected candidates: RMS relative error ε and percentage of data within $\pm 30\%$, both over the filtered database; complexity c = number of nodes.

Rank	id	c	ε	$\pm 30\%$
1	s50-13	13	0.141	95.6%
2	s25-20	20	0.141	95.3%
3	s55-17	17	0.140	95.5%
4	s5-8: Eq. (1)	8	0.170	93.5%
5	s53-20	20	0.156	94.2%

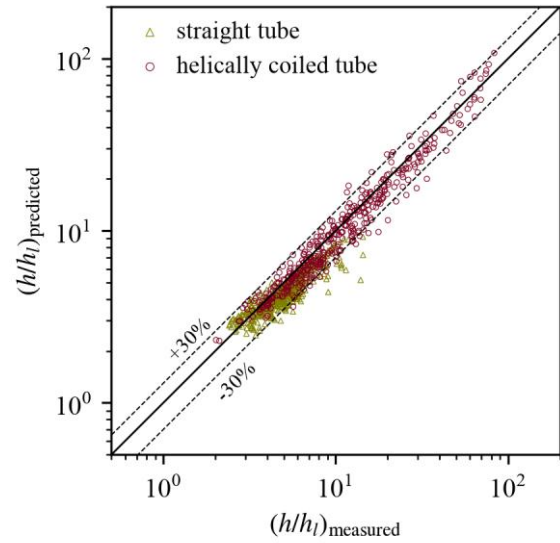


Fig. 1. Comparison of the predicted and measured h/h_i for Eq. (1) over the filtered database.

4. Conclusions

A dimensionally constrained symbolic-regression methodology was presented. A candidate feature set of thermal-hydraulic dimensionless groups ensures that every discovered expression is dimensionally consistent with the target by construction. The exponent of every power is restricted to a numeric constant, and the discovered expressions are collapsed onto common factored forms and screened by their physical behavior. The dimensional constraint is general, whereas the candidate feature set encodes the field-specific physics; other fields can supply their own dimensionless groups. Applied to saturated flow boiling in helically coiled tubes, the methodology reduced the 852 Pareto-optimal expressions from 64 independent runs to the small set listed in Table I. Implementation and evaluation of the selected expressions in a thermal-hydraulic system code remain for future work.

Acknowledgement

This work was supported by the Innovative Small Modular Reactor Development Agency grant funded by the Korea Government(MSIT) (No. RS-2024-00403548).

NOMENCLATURE

Bo	boiling number;
D_c	coil diameter (m);
d	tube inner diameter (m);
G	mass flux (kg/m ² s);
h	two-phase heat transfer coefficient (W/m ² K);
h_l	single-phase heat transfer coefficient (W/m ² K);
P	pressure (Pa);
p_c	coil pitch (m);
Pr	liquid Prandtl number;
q''	heat flux (W/m ²);
Re	Reynolds number of the liquid fraction;
x	quality;
X_{tt}	Lockhart-Martinelli parameter.

REFERENCES

- [1] M. Cranmer, Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl, arXiv:2305.01582, 2023.
- [2] M. G. Kim, B. Yun, and J. J. Jeong, Development of a New Correlation for Saturated Flow Boiling Heat Transfer in a Helically Coiled Tube, Nuclear Technology, Vol. 211, No. 1, pp. 93-110, 2025.
- [3] J. C. Chen, Correlation for Boiling Heat Transfer to Saturated Fluids in Convective Flow, Industrial and Engineering Chemistry Process Design and Development, Vol. 5, No. 3, pp. 322-329, 1966.
- [4] M. M. Shah, Chart Correlation for Saturated Boiling Heat Transfer: Equations and Further Study, ASHRAE Transactions, Vol. 88, No. 1, pp. 185-196, 1982.
- [5] K. E. Gungor and R. H. S. Winterton, A General Correlation for Flow Boiling in Tubes and Annuli, International Journal of Heat and Mass Transfer, Vol. 29, No. 3, pp. 351-358, 1986.
- [6] S. G. Kandlikar, A General Correlation for Saturated Two-Phase Flow Boiling Heat Transfer Inside Horizontal and Vertical Tubes, Journal of Heat Transfer, Vol. 112, No. 1, pp. 219-228, 1990.