Nuclear Energy Agency (NEA)
NEA/WKP(2023)1
January 2024
www.oecd-nea.org

NEA WORKING PAPER

Benchmark on Artificial Intelligence and Machine Learning for Scientific Computing in Nuclear Engineering

Phase 1: Critical Heat Flux Exercise Specifications

Jean-Marie LE CORRE
Gregory DELIPEI
Xu WU
Xingang ZHAO

OECD
BETTER POLICIES FOR BETTER LIVES

NEA
NUCLEAR ENERGY AGENCY

Nuclear Energy Agency

NEA/WKP(2023)1

Unclassified

English text only

17 January 2024

OECD Nuclear Energy Agency
STEERING COMMITTEE FOR NUCLEAR ENERGY

Benchmark on Artificial Intelligence and Machine Learning for Scientific Computing in Nuclear Engineering

Phase 1: Critical Heat Flux Exercise Specifications

A Working Paper by the Task Force on Artificial Intelligence and Machine Learning for Scientific Computing in Nuclear Engineering supervised by the Expert Group on Reactor Systems Multi-Physics (EGMUP)

Jean-Marie LE CORRE, Gregory DELIPEI, Xu WU, Xingang ZHAO

JT03535898

Foreword

Under the guidance of the Nuclear Energy Agency (NEA) Nuclear Science Committee (NSC), the Working Party on Scientific Issues and Uncertainty Analysis of Reactor Systems (WPRS) studies the reactor physics, fuel performance, and radiation transport and shielding in present and future nuclear power systems. In 2022, the WPRS Expert Group on Reactor Systems Multi-Physics (EGMUP) mandated a new Task Force on Artificial Intelligence (AI) and Machine Learning (ML) for Scientific Computing in Nuclear Engineering to develop a benchmark that will provide guidelines and exercises to help participants to develop and evaluate the performance of their artificial intelligence and machine learning methods. The benchmark activity of this Task Force is structured into two phases:

- Phase 1: Regression, Classification and Verification, Validation and Uncertainty Quantification (VVUQ); Dimensionality Reduction and Anomaly Detection.
- Phase 2: Generative Deep Learning and Data Augmentation; Design Optimisation.

This document provides the specifications of the critical heat flux exercise, which is part of the Phase 1 activities.

Acknowledgements

The Nuclear Energy Agency (NEA) is grateful to the participants of the Task Force on Artificial Intelligence and Machine Learning for Scientific Computing in Nuclear Engineering supervised by the Expert Group on Reactor Systems Multi-Physics (EGMUP) for the development of this benchmark specification. Special recognition is extended to the authors and reviewers of this document, whose contributions were instrumental in shaping the benchmark specifications. Additionally, the NEA is profoundly grateful to the US Nuclear Regulatory Commission (NRC) for providing a critical heat flux database and for forming the cornerstone of this benchmark exercise.

Table of contents

Figures

Tables

List of contributors

Authors	**Jean-Marie LE CORRE**	lecorrjm@westinghouse.com	Westinghouse Electric Sweden AB, Sweden
	Gregory DELIPEI	gkdelipe@ncsu.edu	North Carolina State University, US
	Xu WU	xwu27@ncsu.edu	North Carolina State University, US
	Xingang ZHAO	zhaox2@ornl.gov	Oak Ridge National Laboratory, US
Reviewers	**Catalina ANGHEL**	catalina.anghel@cnl.ca	Canadian Nuclear Laboratories, Canada
	Barbara CALGARO	barbara.calgaro@framatome.com	Framatome, France
	Juliana Pacheco DUARTE	pachecoduarte@wisc.edu	University of Wisconsin, US
	Upendra ROHATGI	rohatgi@bnl.gov	Brookhaven National Laboratory, US
NEA Secretariat	**Oliver BUSS**	oliver.buss@oecd-nea.org	OECD Nuclear Energy Agency

List of abbreviations and acronyms

AI	Artificial Intelligence
BFBT	Boiling water reactor (BWR) Full-size Fine-mesh Bundle Tests
BWR	Boiling water reactor
CHF	Critical heat flux
D	Diameter
DL	Deep learning
DNB	Departure from nucleate boiling
EPRI	Electric Power Research Institute (United States)
EGMUP	Expert Group on Reactor Systems Multi-Physics (NEA NSC)
G	Mass flux
H	Enthalpy
L	Heated length
LUT	Lookup table
MAE	Mean absolute error
ML	Machine learning
NEA	Nuclear Energy Agency
NN	Neural network
NRC	Nuclear Regulatory Commission (United States)
NSC	Nuclear Science Committee (NEA)
OECD	Organisation for Economic Co-operation and Development
ONNX	Open neural network exchange
P	Pressure
PCA	Principal component analysis
PSBT	Pressurised water reactor (PWR) Subchannel and Bundle Tests
PWR	Pressurised water reactor
RMSE	Root mean squared error
T	Temperature
UQ	Uncertainty quantification
VVUQ	Verification, Validation and Uncertainty Quantification
WPRS	Working Party on Scientific Issues and Uncertainty Analysis of Reactor Systems (NEA NSC)
X	Thermodynamic equilibrium quality

Executive summary

Recent performance breakthroughs in artificial intelligence (AI) and machine learning (ML), including advances in deep learning (DL) and the availability of powerful, easy-to-use ML toolboxes, have led to unprecedented interest in AI and ML among nuclear engineers. Nonetheless, the extensive capabilities of AI and ML remain largely untapped within the realm of scientific computing in nuclear engineering. One formidable hurdle in harnessing their power is the frequent mismatch between existing ML methodologies and the specific demands of nuclear engineering applications and their extensive validation requirements. To enable more trustworthy applications in high-consequence systems like nuclear reactors that are subject to nuclear safety regulations, the ML practitioners have to address several critical issues, including the verification, validation and uncertainty quantification (VVUQ) of AI and ML, data scarcity, scaling-induced uncertainty, and lack of physics in black-box ML models.

Under the guidance of the NEA Nuclear Science Committee (NSC), the Working Party on Scientific Issues and Uncertainty Analysis of Reactor Systems (WPRS) studies the reactor physics, fuel performance, and radiation transport and shielding in present and future nuclear power systems. In line with the NEA strategic target to contribute to building a solid scientific and technical basis for the development of future generation nuclear systems and the deployment of innovations, the WPRS Expert Group on Reactor Systems Multi-Physics (EGMUP) mandated in 2022 a new Task Force on Artificial Intelligence and Machine Learning for Scientific Computing in Nuclear Engineering to develop and execute benchmarks for AI/ML applications with the following goals:

- to build communities of practice dedicated to the exchange of know-how in the field of AI and ML applications;
- to support the development and performance assessment of ML methods;
- to leverage the insights gained from the benchmarks to distil lessons learnt and to provide guidelines for future AI & ML applications in scientific computing in nuclear engineering.

The benchmark activity of the EGMUP Task Force has been structured into two phases:

- Phase 1: Regression, Classification and VVUQ; Dimensionality Reduction and Anomaly Detection;
- Phase 2: Generative Deep Learning and Data Augmentation; Design Optimisation.

This document provides the specifications of an exercise related to critical heat flux (CHF) that is part of the Phase 1 activities.

In a boiling system, the CHF corresponds to the limit beyond which wall heat transfer decreases significantly. The phenomenon can also be referred to as critical boiling transition (Kaizer et al., March 2019[1]), boiling crisis and (depending on operating conditions) departure from nucleate boiling (DNB), dryout, etc. In a heat transfer controlled system, such as a nuclear reactor core, CHF can result in a significant wall temperature increase, leading to accelerated wall oxidation and potentially fuel rod failure. While constituting an important design limit criterion for the safe operation of reactors, CHF is very challenging to predict accurately due to the complexities of the involved local phenomena. Additionally, large uncertainties are associated with the CHF prediction restricting the reactor design and operation flexibility.

Current CHF models are mainly based on empirical correlations developed and validated for a specific application case domain. Through this benchmark, improvements in the CHF modelling are sought using AI & ML methods directly leveraging the available experimental databases. The improved modelling can lead to a better understanding of the safety margins and provide new opportunities for design or operational optimisations.

Recently, a database used to develop the widely known 2006 Groeneveld CHF lookup table (LUT) was published in digital form by the US Nuclear Regulatory Commission (NRC) (Groeneveld, January, 2019[2]). This database (hereafter referred as the "NRC CHF database"), consisting of nearly 25 000 data points, is the largest known CHF dataset publicly available worldwide with measurements in vertical water-cooled tubes collected over a span of 60 years. Thus, the NRC CHF database provides the opportunity to further develop advanced data-driven regression methods to enable faster and more accurate CHF predictions.

The scope of this report is limited to the specification of Phase 1 of the CHF benchmark exercise organised as part of the Benchmark on Artificial Intelligence and Machine Learning (AI/ML) for Scientific Computing in Nuclear Engineering (NEA, 2023[3]) by the NEA NSC/WPRS/EGMUP Task Force on AI/ML.

In Phase 1 of the CHF exercise, the proposed tasks will leverage the NRC CHF database to train data-driven ML models that could improve the LUT prediction performance. The potential of ML has been demonstrated in an early ML regression analysis of the NRC CHF database using various ML algorithms (Grosfilley, 2022[4]) (Grosfilley et al., 2023[5]).

The benchmark exercises specified in this report define four main AI & ML tasks based on the NRC CHF database. Task 1 is optional and consists of a dimensionality analysis, where the participants are asked to perform feature selection and extraction for their ML models. In Task 2, ML CHF regression algorithms are developed and assessed, including model optimisation, training/validation and testing. In Task 3, the participants evaluate their trained models by computing specific metrics and by ensuring that overfitting does not occur. Finally, in Task 4 CHF predictions on a blind dataset not seen during the training/validation/testing are requested.

The target audience of this report are the participants of the corresponding benchmark activity executed by the EGMUP Task Force on Artificial Intelligence and Machine Learning for Scientific Computing in Nuclear Engineering. Furthermore, the report can serve scientific computing experts as an example of how to benchmark AI/ML applications in nuclear engineering. Even after termination of the benchmark activities by the EGMUP Task Force, the benchmark specifications can serve as input for subsequent benchmark activities addressing future algorithms and exercises. This report is therefore intended to become part of a data package including the results of the benchmark executed by the EGMUP Task Force and linked to the NRC CHF database to serve as valuable input for future benchmark studies.

1. Introduction

In a boiling system, the critical heat flux (CHF) corresponds to the limit beyond which wall heat transfer decreases significantly. The phenomenon can also be referred to as critical boiling transition (Kaizer et al., March 2019[1]), boiling crisis and (depending on operating conditions) departure from nucleate boiling (DNB), dryout, etc. In a heat transfer controlled system, such as a nuclear reactor core, CHF can result in a significant wall temperature increase, leading to accelerated wall oxidation and potentially to heater (e.g. fuel rod) failure. While constituting an important design limit criterion for the safe operation of reactors, CHF is challenging to predict accurately due to the complexities of the involved phenomena, which are not well understood up to date. CHF thus remains a potential source of uncertainties and improvements in the comprehension of CHF dependencies and modelling can directly impact the operational flexibility and safety of nuclear reactors.

The history of CHF predictive model development for convective boiling systems is tightly coupled to the development of civil nuclear power systems. The first measurements date back to 1949 and rapidly expanded in the 1960s and 1970s (Groeneveld, January, 2019[2]). Measurement uncertainties in such experiments can, however, be significant. CHF is known to mainly depend on flow conditions and geometrical parameters. Initial attempts to develop predictive models were based on *empirical correlations*. As interest in CHF prediction increased, various *analytical models* (including mechanistic models) were proposed that could, to some extent, predict CHF across various flow regimes and conditions. However, the level of prediction accuracy desired for reactor design and safety analyses still requires the use of empirical CHF prediction models developed from design-specific data measured over relevant, but limited, operational ranges.

The most successful attempts to correlate CHF over a large parameter space have been performed using *lookup tables*, for which the applicability remains, however, limited. The best-known example of this approach has been documented in (Groeneveld et al., Sept. 2007[6]) and is known as the *2006 Groeneveld CHF lookup table (LUT)*, applicable to an 8 mm (normalised) uniformly heated vertical diameter pipe. Available correction factors can be used to adjust the predictions to various other designs (other diameters, non-uniform power, rod bundle, etc.). The CHF LUT can be considered a data-driven approach using three input parameters (pressure, mass flux and local thermodynamic equilibrium quality) which has the following advantages:

- reasonable accuracy;
- wide range of operating conditions, thus limiting the need for extrapolation;
- the ability to improve predictions by gathering more data.

However, there are still some drawbacks due to data scarcity in some parts of the input domain and incapability to capture complex behaviour stemming from second-order parameters. This results in a relatively large root mean square error (RMSE) of nearly 39% when predicting CHF using constant local conditions (Groeneveld et al., Sept. 2007[6]), making the CHF LUT approach essentially insufficient for many applications.

The CHF database used to develop the 2006 Groeneveld CHF LUT was recently published in digital form by the US Nuclear Regulatory Commission (NRC) (Groeneveld, January, 2019[2]). This database (hereafter referred as the "NRC CHF database"), consisting of nearly 25 000 data points, is the largest known CHF dataset publicly available worldwide with measurements in vertical water-cooled tubes collected over a span of 60 years. Thus,

the NRC CHF database provides the opportunity to further develop advanced data-driven regression methods to enable faster and more accurate CHF predictions.

The scope of this document is limited to the description of Phase 1 of the CHF benchmark exercise organised as part of the Benchmark on Artificial Intelligence and Machine Learning (AI/ML) for Scientific Computing in Nuclear Engineering (NEA, 2023[3]) by the NEA NSC/WPRS/EGMUP Task Force on AI/ML.

In Phase 1 of the CHF exercise, the proposed tasks will leverage the NRC CHF database to train data-driven ML models that could significantly improve upon the LUT prediction performance. The potential of ML has been demonstrated in an early ML regression analysis of the NRC CHF database using various ML algorithms (Grosfilley, 2022[4]) (Grosfilley et al., 2023[5]).

Phase 2 of the CHF exercise (shortly described in Chapter 5 of this report) will further investigate advanced Verification, Validation and Uncertainty Quantification (VVUQ) for the developed ML models and transfer learning for applications to more complex geometries. Eventually, it is envisioned that the ML models would learn the underlying complex dependencies of the CHF phenomena, along with the associated uncertainties, and also be able to acquire knowledge across multiple databases for further improvements. The transfer learning capabilities of the developed ML models will be investigated in order to expand their applicability to domains not covered by the training database.

Chapter 2 of this document provides an overview of the NRC CHF database. Chapter 3 describes the benchmark tasks in detail, while Chapter 4 presents the expected data format to be used by the participants to submit their results. Chapter 5 outlines some initial plans for tasks that will be included in Phase 2 of the CHF exercise. Finally, in Chapter 6, a timeline for the main activities related to the CHF benchmark exercise is provided.

2. Critical heat flux database

The reference experimental database selected for Phase 1 of the CHF exercise is described in Sections 2.1 to 2.3. Additional data, outside this reference database but within the same geometrical and operating ranges, will also be selected by the benchmark organisers for evaluating the participants' ML models on a blind dataset.

2.1. Reference database overview

The NRC CHF database (Groeneveld, January, 2019[2]) is selected as the reference database for Phase 1 of the CHF exercise. The database contains 24 579 CHF measurements in vertical water-cooled uniformly heated tubes compiled from 59 different sources. The available data consists of measured boundary conditions (pressure, P, mass flux, G, inlet temperature, T_{in}, and critical heat flux, CHF), geometrical parameters (test section diameter, D, and heated length, L) and calculated parameters derived from measurements and water properties (outlet equilibrium quality, X, and inlet enthalpy, ΔH_{in}). This database was collected from experimental measurements performed during a span of 60 years, based on various CHF identification methods, such as visual identification, physical burnout, changes in the test section resistances, and the usage of thermocouples.

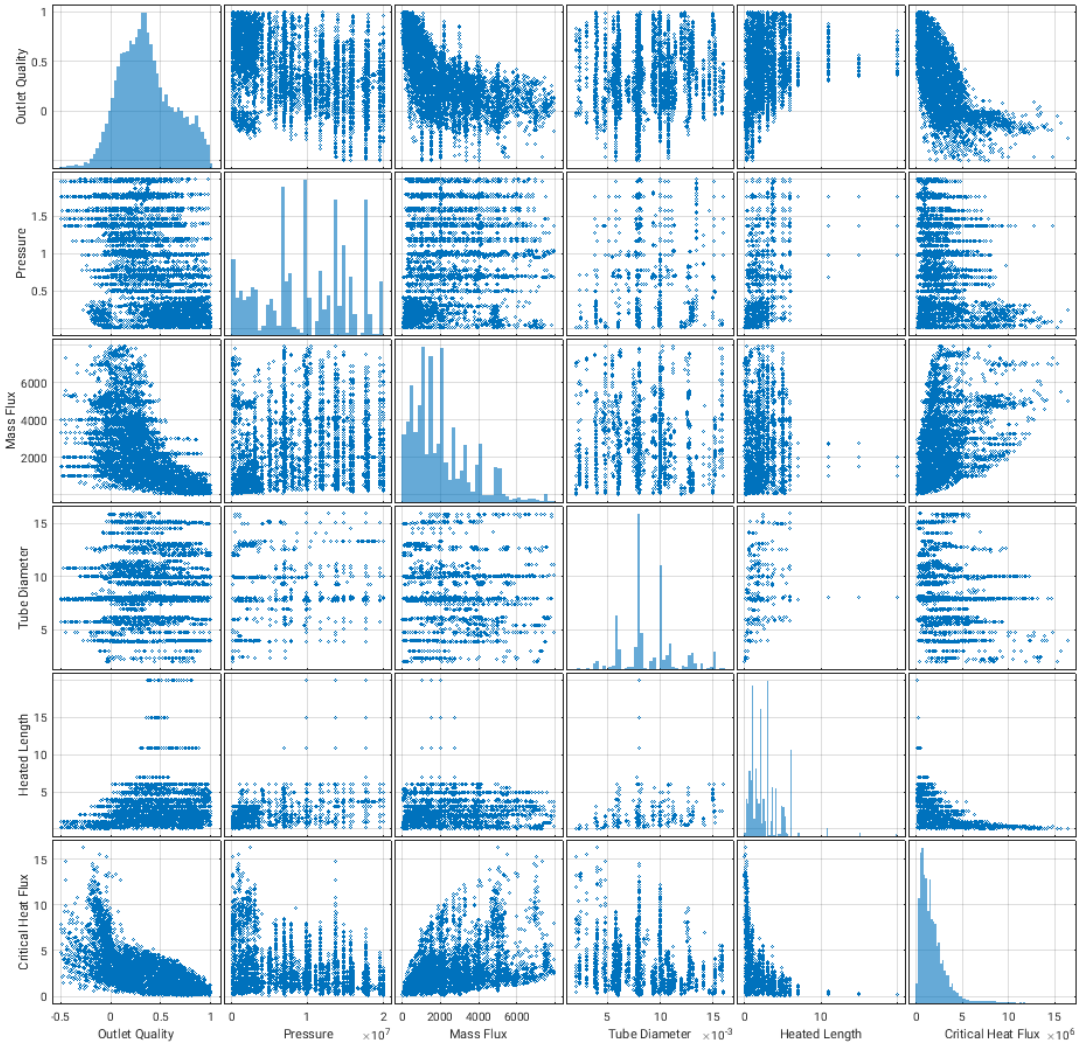
The parameter space covered by the NRC CHF database is substantial, as seen in Table 2.1 and Figure 2.1. In the selection of the data, the database was purposely limited in term of diameter ($2 < D < 25$ mm), L/D ratio ($L/D > 50$ for $X > 0$, $L/D > 25$ for $X < 0$), pressure ($100 \leq P \leq 21\,000$ kPa) and mass flux ($0 \leq G < 8\,000$ kg/m^2/s) (Groeneveld, January, 2019[2]). It should be noted, however, that the measured data is not equally distributed over the whole span and that no data beyond $D = 16$ mm was found in the database.

Table 2.1. Parameter spans of the NRC CHF database

Variable	CHF [kW/m^2]	P [kPa]	G [kg/m^2/s]	X	D [m]	L [m]
Min Value	50	100	8.2	-0.497	$2 \cdot 10^{-3}$	0.05
Max Value	16339.3	20000	7964	0.999	$16 \cdot 10^{-3}$	20

While the NRC CHF database is mostly identical to the database used in the derivation of the 2006 CHF LUT, all proprietary datasets have been removed (Groeneveld, January, 2019[2]). In addition, the database was screened for potential non-physical data, outliers and duplicates (Groeneveld, January, 2019[2]) (Groeneveld et al., Sept. 2007[6]).

Figure 2.1. Scatter plot matrix of the NRC CHF database showing the relationship between pair of variables



2.2. Data subsets and references

All references and an overview of all test facilities that have contributed to develop the NRC CHF database are documented in (Groeneveld, January, 2019[2]). A summary is provided in Table 2.2, including the number of data points and the parameter ranges covered by each considered dataset in terms of tube diameter, heated length, pressure, mass flux, outlet quality and inlet temperature. Note that this summary table is slightly different from Table 4-2 of (Groeneveld, January, 2019[2]), which includes additional data, such as proprietary data and data not included in the development of the 2006 CHF LUT.

For the proposed CHF exercise, a reference ID (included in Table 2.2) was assigned to each dataset according to the order encountered in the data file.

An effort will be made to retrieve and share the original references from which the NRC CHF database was developed. To support this effort, participants having access to some of the original references are encouraged to contact the benchmark organisers.

Table 2.2. Datasets and parameter ranges (min–max) overview of the NRC CHF database

ID	Reference from [2]	Number	Tube diameter mm		Heated length m		Pressure MPa		Mass flux kg/m^2/s		Outlet quality -		Inlet temperature C	
13	Alekseev 1964	1057	10.0	10.0	1.00	4.97	9.800	19.610	216	7566	-0.311	0.944	91.8	350.1
7	Alessandrini et al 1963	161	15.2	15.2	0.80	2.46	4.795	5.109	1100	4140	-0.040	0.740	185.6	265.3
28	Babarin et al 1969	103	12.0	12.0	0.96	1.80	0.290	0.310	100	500	0.510	0.997	16.0	124.0
3	Babcock and Hood 1962	11	8.0	14.2	0.61	0.61	0.413	6.890	4101	7302	-0.159	-0.050	19.4	191.5
58	Baek 2001 from KAIST	34	6.0	10.0	0.18	0.40	0.101	3.618	803	2032	-0.091	0.099	5.9	40.2
38	Bailey 1977	113	15.0	15.0	3.77	5.37	1.350	7.080	49	1383	0.450	0.990	92.9	286.3
29	Bailey and Lee 1969	157	9.3	9.3	3.05	3.05	6.895	18.340	958	4242	0.069	0.727	198.7	346.9
9	Becker 1963 AE 114, 1965 AE-177	809	10.0	10.0	0.60	2.50	0.520	4.884	100	2515	0.204	0.996	30.2	172.8
8	Becker 1965 AE-177 Table 1.1	749	3.9	10.0	0.96	3.12	0.235	5.776	153	2410	0.181	0.999	16.0	187.1
10	Becker 1965 AE-177 Table 1.2	752	10.0	13.1	0.40	3.00	0.216	3.844	111	1565	-0.069	0.909	16.1	147.7
11	Becker 1965 AE-177 Table 1.3	169	6.1	10.0	1.00	2.00	0.981	4.021	168	3183	0.096	0.907	30.1	63.1
17	Becker et al 1965 AE-178, AE-177	661	3.9	10.0	1.00	3.50	1.128	9.905	222	5451	-0.005	0.993	37.6	229.9
32	Becker et al 1970	69	2.4	3.0	0.50	0.50	3.100	7.100	365	2725	0.207	0.891	29.1	77.7
34	Becker et al 1971	1435	10.0	10.0	1.00	4.97	3.000	20.000	156	7568	-0.311	0.997	86.9	357.4
18	Bennett 1965 AERE R5055	198	9.2	12.6	1.73	5.56	6.612	7.481	624	5844	0.026	0.948	135.6	279.5
12	Bergles 1963	7	2.4	2.4	0.06	0.08	0.207	0.207	3037	6075	-0.037	-0.035	14.6	50.7
19	Burch and Hufschmidt 1965	134	10.0	10.0	0.35	0.35	1.100	3.090	930	3756	-0.246	0.000	16.8	60.3
50	Celata 1992 Revue Thermique	7	2.5	5.0	0.20	0.40	0.107	2.116	2166	5905	-0.015	0.288	18.9	22.5
51	Celata and Mariani 1993	7	4.0	4.0	0.10	0.10	0.794	2.508	4924	5157	-0.276	-0.107	30.2	69.8
42	Cheng et al 1983a 1983b	116	4.8	12.3	0.39	0.74	0.100	0.700	50	750	0.187	0.998	49.6	154.5
30	Dell et al. 1969	82	6.2	6.2	0.91	5.51	6.895	6.895	1329	4136	0.144	0.779	209.5	269.7
26	Era et al 1966	151	6.0	6.0	1.60	4.80	6.777	7.049	1105	3006	0.374	0.952	165.2	286.3
40	From Kirillov's data base 1992	271	8.0	8.0	0.24	0.40	0.170	3.080	1999	7078	-0.217	-0.001	27.4	148.2
20	Griffel 1965	218	6.2	12.8	0.91	1.93	5.171	10.343	936	7783	-0.069	0.592	45.9	285.5
44	Groeneveld 1985	117	10.0	10.0	1.00	2.00	7.900	20.000	282	2805	-0.097	0.805	76.9	220.4
21	Hewitt 1965	289	9.3	9.3	0.61	3.05	0.110	0.208	91	301	0.462	0.997	13.4	121.3
6	Hood and Isakoff 1962	10	8.0	12.9	0.60	1.11	6.895	6.895	664	2007	0.020	0.419	7.5	194.8
5	Hood 1962	16	8.0	14.2	0.61	0.61	0.414	8.412	2753	7200	-0.239	-0.052	17.5	189.9
48	Inasaka and Nariai 1989	3	3.0	3.0	0.10	0.10	0.310	0.910	5500	6700	-0.115	-0.056	26.0	54.0
52	Jafri 1993	12	15.7	15.7	2.44	2.44	0.362	1.060	1456	7830	0.095	0.435	74.4	171.5
33	Jens and Lottes 1951	29	5.7	5.7	0.63	0.63	3.448	13.790	1302	5356	-0.464	-0.021	48.7	266.9
24	Judd and Wilson 1966	49	11.3	11.3	1.83	1.83	6.861	13.859	674	3428	0.016	0.776	193.8	322.2
57	Kim et al 2000	482	6.0	12.0	0.30	1.77	0.104	0.951	40	277	0.458	0.999	20.5	156.3
43	Kirillov 1984 1985	2401	7.7	8.1	0.99	6.00	6.370	18.040	494	4154	-0.494	0.981	34.7	344.7
56	Kureta 1997	140	2.0	6.0	0.05	0.68	0.101	0.101	8	7156	-0.064	0.991	7.8	32.7
25	Lee 1966 AEEW R479	257	14.1	14.1	0.64	1.52	8.616	12.476	530	3410	-0.078	0.523	233.8	317.7
47	Leung et al 1989	62	5.5	5.5	2.51	2.51	5.030	9.710	1168	7442	0.210	0.578	221.9	305.1
1	Lowdermilk 1958	61	4.0	4.8	0.40	0.99	0.100	0.100	69	1645	0.420	0.990	20.6	23.9
22	Matzner et al 1965	83	10.2	10.2	2.44	4.88	6.893	6.893	1193	7960	0.008	0.693	17.3	275.7
27	Mayinger et al 1966	102	7.0	7.0	0.56	0.98	1.925	10.244	2255	3578	0.098	0.353	159.3	312.3
45	Nariai et al 1987	7	2.0	3.0	0.05	0.10	0.100	0.100	6900	7350	-0.070	-0.016	20.5	60.7
37	Nguyen and Yin 1975	56	12.6	12.6	2.44	4.88	6.645	8.401	930	3838	0.216	0.738	215.2	276.7
49	Olekhnovitch 1991, 1997, 1999	194	8.0	8.0	0.75	3.50	0.525	4.007	988	6082	0.046	0.761	59.7	244.5
55	Pabisz and Bergles 1966	6	4.4	4.4	0.11	0.11	0.872	1.284	3801	4567	-0.196	-0.147	22.1	46.2
14	Peterlongo et al 1966	342	15.1	15.2	2.24	4.02	4.943	6.551	1010	4020	-0.023	0.608	27.1	281.1
59	Shan 2004	70	8.0	15.8	1.00	2.44	0.317	14.808	572	7830	-0.022	0.422	18.9	328.4
2	Smolin 1962	616	3.8	10.8	0.78	4.00	7.840	19.610	498	7556	-0.132	0.786	64.7	346.5
39	Smolin 1979	2928	3.8	16.0	0.69	6.05	2.940	17.710	490	7672	-0.136	0.789	46.5	349.3
53	Soderquist 1994	1250	8.0	8.1	1.00	6.00	0.970	20.000	246	6086	-0.043	0.999	111.6	354.6
4	Swenson 1962	25	10.4	10.5	1.75	1.80	13.790	13.790	679	1765	0.178	0.502	231.4	329.4
54	Tain 1994	55	8.0	8.0	1.75	1.75	6.849	10.127	2401	7832	0.028	0.378	191.5	299.1
15	Tong 1994	218	6.2	12.9	0.76	3.66	5.171	13.790	678	7960	0.002	0.502	46.9	330.4
0	Unknown sources	384	5.6	12.8	0.43	2.01	5.270	7.490	405	5586	0.000	0.950	76.8	286.1
23	Waters et al 1965	17	11.2	11.2	1.52	3.65	6.895	10.342	6578	7961	-0.034	0.275	86.9	313.5
41	Williams and Beus 1980	128	9.5	9.5	1.84	1.84	2.758	15.169	324	4663	-0.025	0.929	90.6	315.6
46	Yin et al 1988	250	13.4	13.4	3.66	3.66	1.028	19.989	1939	2082	0.164	0.431	126.7	357.0
35	Zenkevich 1971	361	7.8	8.1	7.00	20.00	6.860	17.650	1008	2783	0.262	0.876	36.3	351.8
36	Zenkevich 1974	835	4.8	12.6	1.00	6.00	5.890	19.620	497	6694	-0.221	0.969	35.1	357.4
16	Zenkevich et al 1964	1	8.0	8.0	0.20	0.20	3.924	3.924	5361	5361	-0.004	-0.004	212.0	212.0
31	Zenkevich	5252	4.0	15.1	0.25	6.00	5.880	19.610	498	7964	-0.497	0.964	22.7	361.9
Total		**24579**	**2.0**	**16.0**	**0.05**	**20.00**	**0.100**	**20.000**	**8**	**7964**	**-0.497**	**0.999**	**5.9**	**361.9**

2.3. Experimental data format

The NRC CHF database is provided to the participants in the file `chf_public.csv` using a comma-separated values (csv) data format. This file consists of a cleaned-up version of the document obtained from the US NRC (Groeneveld, January, 2019[2]) with a few formatting changes to facilitate access using computer scripts. An example (when opened in Excel) of the first five data points can be seen in Table 2.3.

Table 2.3. Example of data format used for distribution of the NRC CHF database

Number	Reference ID	Tube diameter	Heated length	Pressure	Mass flux	Outlet quality	Inlet subcooling	Inlet temperature	CHF
		m	m	kPa	kg/m^2/s		kJ/kg	^{0}C	kW/m^2
1	1	0.004	0.396	100	77.5	0.84	317	23.94	442
2	1	0.004	0.396	100	142.7	0.79	317	23.94	757
3	1	0.004	0.396	100	203.9	0.7	317	23.94	978
4	1	0.004	0.396	100	271.8	0.73	317	23.94	1 325
5	1	0.004	0.396	100	421.3	0.62	317	23.94	1 798

The provided data can be classified into three categories: (1) geometry (tube diameter and heated length), (2) measured parameters (pressure, mass flux, inlet temperature and CHF), and (3) calculated parameters (outlet quality and inlet subcooling). The reference ID refers to the test facility ID listed in Table 2.2.

The calculated input parameters (outlet quality and inlet subcooling) are derived from the measured parameters and water properties at saturation. For instance, the outlet equilibrium quality can be calculated from inlet temperature, power, mass flow and pressure-dependent fluid properties, based on energy conservation. These calculated values originate from the database provided by the US NRC (Groeneveld, January, 2019[2]). As needed, the participants can recalculate these parameters using the water properties of their choice (small variations may be observed). Note that these calculated input parameters provide *redundant* information that requires careful considerations when selecting the independent parameters (or features) of any ML regression model (see further discussion in Section 3.1).

2.4. Test matrix for the benchmark activities

The considered databases for Phase 1 of the CHF exercise are divided into (1) training/validation/testing datasets, (2) blind test dataset and (3) "slice" datasets.

2.4.1. Training/validation/testing datasets

The entire NRC CHF database may be considered for training, validation and testing. The participants are free to use any amount of data points from the provided database, any ML techniques to build the regression model, and any approaches for training and validation to improve the ML models' prediction accuracy. The training, validation and testing datasets follow the definitions provided in Appendix A. Common techniques are hold-out tests and *k*-fold cross-validation. The purpose of these datasets is to support the selection, training and tuning of the final ML models that will eventually be tested with the unseen blind test dataset.

2.4.2. Blind test dataset

The ML models developed by the participants will also be assessed using a separate blind tests dataset. This dataset will be selected outside the NRC CHF database and known only by the benchmark organisers. The same generic geometrical and test characteristics will remain, that is, vertical uniformly heated tube geometries within the parameter space covered by the NRC CHF database (see Table 2.1).

As a separate assessment, keeping in mind the limitations in extrapolation of all ML models techniques, a dataset outside this parameter space could also be selected to analyse CHF prediction models performances in such conditions.

2.4.3. "Slice" datasets

In data analysis, slicing methods allow reducing databases into smaller and coherent parts, allowing extraction of further information and deeper understanding. An example of slicing method was used in (Groeneveld et al., Sept. 2007[6]) to identify experimental outliers and analyse how CHF varies as equilibrium quality increases (everything else being constant).

In preparation for this benchmark, a similar and systematic slicing of the NRC CHF database has been performed where interesting regions of varying D, L, P, G and X have been identified, with all other parameters remaining reasonably constant. Among the identified data slices, two slices per varying parameters have been selected and documented in Table 2.4. In the provided databases (with file names identified in the table), the varying parameters are arbitrarily divided into 15 equidistant values, across the ranges defined in Table 2.1. These "slice" datasets will be used by the participants to demonstrate the physical behaviour of their predictive CHF models across the considered parameter space, including the prevention of overfitting. This will facilitate the explainability of the models and provide more confidence in their predictive capabilities.

Table 2.4. Slice datasets

Slice #	D [mm]	L [m]	P [kPa]	G [kg/m^2/s]	X [-]	Data file
1	**0 - 16**	6.000	14701	998.5	0.391	Slice_01.csv
2	**0 - 16**	6.000	9807	1003.3	0.529	Slice_02.csv
3	8.01	**0 - 20**	9806	1000.0	0.587	Slice_03.csv
4	8.11	**0 - 20**	2009	752.2	0.756	Slice_04.csv
5	8.00	0.998	**0 - 20000**	2006.0	0.140	Slice_05.csv
6	13.40	3.658	**0 - 20000**	20040.2	0.378	Slice_06.csv
7	8.00	1.570	12750	**0 - 8000**	0.144	Slice_07.csv
8	10.00	4.966	16000	**0 - 8000**	0.343	Slice_08.csv
9	8.14	1.943	9831	1519.5	**-0.5 - 1.0**	Slice_09.csv
10	8.00	0.997	17650	2002.7	**-0.5 - 1.0**	Slice_10.csv

It can be noted that slice 9 includes a significant decrease in CHF with increasing quality at around $X = 0.4$, which is referred to as the "limiting quality region" in (Groeneveld et al., Sept. 2007[6]).

3. Benchmark tasks

For the CHF exercise, the optional activities related to dimensionality analysis (Section 3.1) will be performed before proceeding to the regression task (Section 3.2) and VVUQ (Section 3.3). The requested data format for each task is documented in Chapter 4. Further activities, including advanced UQ and transfer learning to complex geometries using ML tools will be addressed subsequently and proposed in Phase 2 of the CHF exercise (see Chapter 5).

3.1. Dimensionality analysis (Task 1) - optional

CHF regression is commonly performed using predictive multivariate models, based on several input thermal-hydraulic quantities. Currently, most CHF prediction models for convective boiling systems use analytical functions that mainly depend on pressure, *P*, local mass flux, *G*, channel diameter, *D*, and local equilibrium quality, *X*:

$$\mathrm{CHF} = f(P, G, D, X)$$

Additional parameters are sometimes also included (such as the heated length, *L*). Other alternatives to this set of input parameters based on saturated fluid properties (instead of *P*), or non-dimensional formulations can also be found (e.g. in (Hall and Mudawar, 2000[7])), which have the advantage to be potentially applicable to different fluids.

The NRC CHF database offers the opportunity to perform dimensionality analysis with the aim to (1) reduce and (2) select the best performing input/output parameters for CHF model predictions. This analysis can be supported by ML dimensionality reduction techniques, which can be divided into (1) feature selection and (2) feature extraction.

Applying these techniques has also the potential advantages of improving data scaling, speeding up training, lowering the risk of overfitting, and increasing the explainability of the model.

This benchmark activity is optional but highly recommended. The participants may use any type of AI/ML methodology to select the most relevant input/output parameters for CHF regression based on the NRC CHF database and any other considerations (e.g. physical principles). In general, any of the available inputs or features that can be constructed from the available data could be used (accounting for the recommendations provided in this section).

3.1.1. Feature selection (Task 1.1)

Feature selection techniques allow reducing the number of variables, and avoiding redundancy in the process of developing a predictive model. Redundant variables tend to reduce the model's generalisation capability and increase the overall complexity of the model. Various feature extraction techniques are available such as filter, wrapper and intrinsic methods.

The participants may use any method of their choice, in combination with the NRC CHF database, to support this analysis. Note that, as a starting point, redundant parameters included in the data file have already been discussed in Section 2.3.

For this activity, the following recommendations are provided to the participants:

- Attention should be given when selecting the set of input parameters to ensure that they are truly independent from CHF. In particular, when selecting the local equilibrium quality as input (calculated from other inputs parameters, including CHF), at least one of the parameters used in the heat balance equation (beside the heat flux) must be left out.
- Correlation (or sensitivity) does not mean causation. In other words, an observed CHF sensitivity with an input variable might not be physically relevant. For instance, a significant but unexpected dependency of CHF with heated length was found in (Grosfilley, 2022[4]) and (Grosfilley et al., 2023[5]) using the NRC CHF database. Such dependency cannot, however, be reasonably justified for large L/D as discussed in (Groeneveld et al., Sept. 2007[6]). Hence other, more relevant, parameters must be found to adequately capture this dependency. The same reasoning would also apply to any inlet-specific dependency, such as T_{in}, at large L/D.

3.1.2. Feature extraction (Task 1.2)

Feature extraction techniques allow combining model variables into new, more relevant, variables (of same or lower dimensions). For instance, the transformation of a model from a dimensional to a non-dimensional form can be considered a type of feature extraction (however not data-driven). Feature extraction can also help in data scaling in a more appropriate way than model training. Various feature extraction techniques, such as principal component analysis (PCA), are relevant and can be considered by the participants.

Domain knowledge based feature engineering, which is closely related to feature extraction, could also be considered. The added value of the additional features should be ensured with respect to the explainability of the model.

3.2. Machine learning regression (Task 2)

The goal of this activity is to develop and assess data-driven (possibly, but not necessarily, physics-informed) AI/ML supervised regression algorithms to predict CHF using the set of input/output parameters selected from the outcome of the dimensionality analysis (see Section 3.1). While of interest, fully physics-based model development is not relevant to this benchmark. Participants that do not perform the dimensionality analysis can use the parameters of the equation provided in Section 3.1 (or any other well-defined formulation) as input/output parameters to the regression task.

3.2.1. Model optimisation (Task 2.1)

Any ML regression algorithms (and any variations within the same algorithm) can be selected and investigated for this activity. In this task, various optimisations of the selected algorithms with respect to data scaling, regularisation and all relevant hyperparameters will be assessed by the participants.

- ML algorithms are generally dependent upon data scaling. For instance, when used without the consideration of feature extraction (Section 3.1.2), the input features yield order of magnitude differences over widely varying spans (as seen in Table 2.1). Data normalisation is hence recommended to utilise the database most efficiently across the parameter space. Common methods are based on Z (standardisation) or Min-Max normalisation.

- The use of regularisation methods is recommended to develop a well-behaved model, which is able to generalise without overfitting. Methods such as L_2 regularisation or dropout may be used.
- ML model hyperparameters tuning can be performed manually, or automatically, using widely available tools such as Optuna (Preferred Networks, Inc, 2023[8]). The models can be optimised with respect to performance (e.g. low loss function) together with any other relevant target (e.g. model size) by adjusting the model's hyperparameters. For example, in the case of a neural network: the network architecture (e.g. number of hidden layers, number of neurons in each layer), the activation/loss functions and all parameters used for training (e.g. learning rate, number of batches, batch size, dropout rate) can potentially be optimised.

The candidates can then report and select their best performing model(s) and proceed with final model training, validation (if applicable), and testing.

3.2.2. Model training and validation (Task 2.2)

After optimisation, the final training and validation of the selected ML predictive model(s) may be performed using the dataset and validation methods discussed in Section 2.4.1. Various outputs from the model training and validation will be requested, as documented in Section 4.2.2.

3.2.3. Model testing (Task 2.3)

A best practice to assess the prediction performances of a ML model (in particular with respect to overfitting) consists in testing the predictions using data unseen by the training and validation algorithms. While a separate testing evaluation will be performed by the benchmark organisers using a dataset outside the NRC CHF database (see Section 3.4), it is recommended for the participants to perform their own model performance assessment (using part of the NRC CHF database or any other datasets). If performed, outputs from this model testing will be requested, as documented in Sections 4.2.3 and 4.3.1.

Eventually, the selected ML models will be applied over the entire NRC CHF database and the participants will be requested to provide the corresponding CHF predictions.

3.3. Model evaluation (Task 3)

Following training/validation and testing, the predictive ML model(s) selected by the participants in this task should be further evaluated with the objective to (1) assess the overall performance and (2) demonstrate that the model(s) are not overfitting. Note that in all activities, the predictions of CHF must be reported based on constant local conditions (Groeneveld et al., Sept. 2007[6]), that is, directly based on inputs specified in the NRC CHF database.

Examples of requested evaluations can be found in Appendices C and D of this document, respectively using the CHF LUT (Groeneveld et al., Sept. 2007[6]) and the neural network (NN) documented in (Grosfilley, 2022[4]) and (Grosfilley et al., 2023[5]) as predictive models.

3.3.1. Model performance assessment (Task 3.1)

The participants should perform a model evaluation over the considered datasets described in Sections 2.4.1 (training/validation), 2.4.2 (testing) and the entire NRC CHF database using the metrics defined in Appendix A in terms of predicted/measured CHF, that is:

mean, standard deviation, root mean squared error (RMSE) in percentage error, mean absolute error (MAE) in percentage error, EQ^2 and provide the results to the benchmark team using the corresponding templates (see Section 4.3.1).

In addition, various CHF prediction assessment plots should be generated and provided, following the examples provided in Appendices C and D:

- measured vs predicted CHF;
- predicted/measured CHF histogram;
- predicted/measured CHF scatter plots vs mass flux, outlet quality, pressure, diameter, heated length and heated length/diameter.

Note that for model comparison purposes, the benchmark organisers will directly use the tabulated model outputs provided by the participants (see Section 4.3.1).

3.3.2. Model behaviour and overfitting assessment (Task 3.2)

The participants should ensure that the selected regression algorithm(s), and associated hyperparameters and training methodology, can generate a well-behaved CHF prediction model with no tendency to overfit. For that purpose, a task based on selected "slice" datasets (as documented in Section 2.4.3) has been defined. The participants should submit the prediction outputs of their CHF model(s) for the ten sets of conditions ("slices") defined in the data files provided in Table 2.4.

Examples of such file with varying parameter *D* (everything being held constant) and data plots for all slices are shown in Appendices C and D, along with CHF predictions from the considered predictive models. The models are expected to yield physical behaviour across the investigated parameter range, following the trend shown by the available data for the selected conditions.

3.4. Independent model evaluation (Task 4)

The benchmark team will gather and compile all the submitted results to perform data analysis and uncertainty assessment, similar to Task 3.1 performed by the participants. The various models will be cross-compared using the predictive metrics documented in Appendix A and trends across models will be computed. Note that a subgroup of the EGMUP Task Force on AI/ML is expected to be created to define acceptability criteria for AI/ML models and methods to assess that AI/ML models perform well for a specific safety related task. Based on recommendations from this subgroup, the benchmark team will consider performing additional model evaluations.

In addition, an independent performance assessment of the participants' ML models will be conducted by the benchmark organisers using a blind test dataset, which will be selected according to the criteria described in Section 2.4.2. The blind dataset provided to the participants will consist of the same input parameters as for the NRC CHF database, except for the CHF, the inlet temperature, and the inlet subcooling. Since the local quality will be provided, the inlet thermal conditions must be withdrawn as a simple heat balance would otherwise provide the corresponding heat flux (note that, following the discussion in Section 3.1.1, the use of inlet properties as input parameter is anyway not recommended). The blind test dataset will be made available to the participants under the file name `chf_blind.csv`.

The participants will provide the CHF predicted by their trained ML model(s) for all conditions within the blind test dataset. The benchmark organisers will then compile the

results and perform cross-comparisons between the various AI/ML models and the actual measured CHF values. Participants are free to submit the CHF predictions for more than one ML model, as long as they follow the benchmark guidelines for each model. When submitting more than one model, however, participants should rank them by order of expected performance.

4. Submission data (Phase 1)

The following sections outline the data and information required from each participant for Phase 1 of the CHF exercise. The submission itself will be handled via the NEA GitLab system and technical information on the submission process is provided on GitLab[1]: https://git.oecd-nea.org/science/wprs/egmup-task-force-on-ai-and-machine-learning/phase1-chf-exercises/-/blob/main/README.md.

4.1. Dimensionality analysis (Task 1)

Note that this task is optional.

4.1.1. Feature selection (Task 1.1)

This activity is documented in Section 3.1.1. The participants will provide the considered variables (among D, L, P, G, T_{in} or ΔH_{in} and X) and all relevant information supporting this selection.

4.1.2. Feature extraction (Task 1.2)

This activity is documented in Section 3.1.2. The participants will provide the considered input and output parameters of their model(s) and all relevant information supporting this selection.

4.2. Machine learning regression (Task 2)

4.2.1. Model optimisation (Task 2.1)

Participants are responsible for selecting their regression ML algorithm and reporting its architecture. The approach used to optimise the model's hyperparameters should be reported in detail. This includes the hyperparameter optimisation framework (if any), optimisation target(s) and the resulting algorithm architecture, data scaling and all training parameters (e.g. regularisation method, loss function, batch size, learning rate, activation function).

It is important that any method employed is documented and that the results are properly reported. An example of submission template is provided in `model_summary.xlsx` and shown in Table 4.1, which includes the hyperparameters related to a NN algorithm. For other considered algorithms, the participants should include all relevant hyperparameters. In the case where the combined output from several models is used as an ensemble, the description of each model should be provided, as well as how the outputs were combined.

In addition, all modules and libraries required by the model, including version number, should be listed.

4.2.2. Model training and validation (Task 2.2)

Participants will perform the training and validation of their selected ML model(s). The following information should be reported:

[1] Access to NEA GitLab is restricted to registered benchmark participants.

- the training/validation data selection and methods;
- the loss curve (training and validation loss vs epoch number), if applicable;
- the trained parameters associated with the selected algorithm(s) (e.g. weights and biases in the case of a neural network) to ensure transparency and reproducibility;
- the trained model(s) in an open format, such as Open Neural Network Exchange (ONNX, see https://onnx.ai/).

4.2.3. Model testing (Task 2.3)

The test data selection method should be reported.

Table 4.1. Template of model descriptions, associated hyperparameters and evaluation results (example for a NN model)

General		Algorithm	Neural network
		Implementation platform	
		Hyperparameter optimization	
		Hyperparameter opt. target	
		Inputs	
		Output	
		Layer architecture	
		Activation function	
Training	Hyperparameters	Data scaling	
		Batch size	
		Training steps per epoch	
		Optimizer	
		Regularization	
		Learning rate	
		Learning rate decay factor	
		Loss function	
		Validation	
		Early stopping	
		Number of epochs	
		Size of data set	
	Evaluation	Mean P/M	
		Std P/M	
		RMSPE [%]	
		MAPE [%]	
		Q^2 error	
Test		Size of data set	
	Evaluation	Mean P/M	
		Std P/M	
		RMSPE [%]	
		MAPE [%]	
		Q^2 error	
All		Size of data set	
	Evaluation	Mean P/M	
		Std P/M	
		RMSPE [%]	
		MAPE [%]	
		Q^2 error	

4.3. Model evaluation (Task 3)

4.3.1. Model performance assessment (Task 3.1)

The participants should provide the CHF predicted by their model for the 24 579 data points of the NRC CHF database. The requested data format is the same as the database format (`chf_public.csv`), with one additional column for the predicted CHF. Two examples of the results files can be found in Table C.2 for the CHF LUT predictions and in Table D.2 for the CHF predictions from the NN documented in (Grosfilley, 2022[4]) and (Grosfilley et al., 2023[5]).

The model performance assessment will be performed by the participants and reported using the template `model_summary.xlsx` shown in Table 4.1, using the metrics defined in Section 3.3.1. The training/validation dataset, test dataset and the entire NRC CHF database will be considered for this evaluation. An example of such assessment can be found in `model_summary_NN.xlsx` and in Table D.1 [for the CHF predictions from the NN documented in (Grosfilley, 2022[4]) and (Grosfilley et al., 2023[5])].

Finally, participants will report all figures and associated values described in Section 3.3.1. The example of such figures can be found in Appendix C (for the LUT) and Appendix D [for the NN documented in (Grosfilley, 2022[4]) and (Grosfilley et al., 2023[5])].

4.3.2. Model behaviour and overfitting assessment (Task 3.2)

The participants should provide the results of their CHF model for the 10 data slices defined in Section 3.3.2. The boundary conditions for the data slices are provided in files `Slice_XX.csv` where `XX` varies from `01` to `10` (see Table 2.4). The CHF predicted by the model(s) will be provided by the participants in an additional column following the same data format. The example of such data file using the CHF LUT results can be found in Table C.4. Similar examples for the NN [documented in (Grosfilley, 2022[4]) and (Grosfilley et al., 2023[5])] can be found in Table D.3.

As part of this task, participants should ensure the reasonable behaviour of their model across the investigated parameter space. For this purpose, the predicted CHF can be compared against the experimental data at similar conditions and against the LUT predictions, both provided in Appendix C.

4.4. Independent model evaluation (Task 4)

Based on the tabulated inputs from the participants, the benchmark team will generate similar outputs and plots as requested in Section 4.3 for each model and compile the models.

In addition, the results from the blind test database will be compiled and compared against the experimental data.

All results and conclusions will be provided to the benchmark participants and documented in a report.

5. Phase 2 Initial plans

After completion of the Phase 1 activities, it is envisioned that participants proceed with the Phase 2 CHF benchmark exercise. Phase 2 will include more advanced model developments and evaluations, as outlined in this section. Modifications and additional analyses can also be considered, based on the feedback from the participants and the EGMUP Task Force on AI/ML.

5.1. Advanced uncertainty quantification

Consistency of model performances across the database will be performed (for instance, using ANOVA method). However, such approach can be superseded by various methodologies related to advanced uncertainty quantification (UQ) of ML regression algorithms proposed in the AI community and still under development. For example, dropout-based, ensemble-based and Bayesian neural network-based approaches for UQ of deep NNs. In essence, these methodologies would allow the quantification of uncertainties associated with any CHF prediction, supported by the prediction error with respect to available data in the considered region (per contrast to a "global" uncertainty quantification applicable across the entire validation range). Both epistemic and aleatory uncertainties are of interest, representative of model and measurement uncertainties, respectively.

For the next phase of the benchmark, a task will be proposed where participants will select the UQ methodology of their choice and report their results.

5.2. Transfer learning to other geometries

Transfer learning has been widely used in the ML field, where various techniques leverage previously acquired data-driven knowledge to slightly different applications. When using transfer learning, knowledge gained from a database is transferred from one "region" of applications (e.g. CHF in simple geometry) to another (e.g. CHF in complex geometry). The purpose is to apply underlying physics learnt from a given database to a "region" with similar underlying physical behaviour but originates from another (potentially more complex) system (e.g. CHF in tubes vs. CHF in subchannels of a rod bundle).

As an example, transfer learning in NNs can be performed by freezing a trained base model and adding new layers before the output layer. The added layers are then trained on a different dataset (e.g. representative of different geometry), hence retaining the knowledge of the base model.

For the next phase of the benchmark, a CHF database from a geometry other than a tube (e.g. an annulus) will be selected. The participants will use their trained model with the NRC CHF database as a base model and transfer it to the new geometry. This task will be developed in several steps, with increasing complexity (including, to the extent possible, the effects of flow mixer, non-uniform axial power on CHF) up to realistic CHF predictions in fuel bundles.

5.3. Model interpretability and explainability

Various techniques can be employed in support of "interpretation" and "explanation" of ML models. This can be useful to avoid a full "black box" approach, which is desirable in particular for regulated industries. "The degree of *interpretability* of a ML model refers to

the degree of understanding of how a ML model is generating its predictions, i.e. to " answer the exact *why* and *how* of the model's behaviour" (Amazon Web Services, 2023[9]). *Explainability* is a limited concept compared to *interpretability*, and refers to the users' capability to grasp how model input influences model output."

Interpretability is strongly related to the model itself. For instance, a decision tree algorithm tends to be highly interpretable while a neural network is typically poorly interpretable. Explainability, on the other hand, is related to the model behaviour, which can be assessed using various methods.

A model interpretability and explainability task will be defined, after further interactions between the participants and the EGMUP Task Force on AI and Machine Learning. One of the main goals of the task will be to provide relevant methods and guidelines in support of the licensing of CHF regression models based on ML algorithms.

5.4. Fuel bundle benchmark

For realistic applications to nuclear power plant safety analysis, ML CHF regression algorithms must be developed for fuel bundle geometries, accounting for the results of operating conditions, geometry (including spacer grids) and three-dimensional power distribution.

As a final step for the CHF exercise, it is planned to organise a benchmark against a large CHF database in fuel bundles. The CHF database will be generated from publicly available datasets such as the Electric Power Research Institute (EPRI), (EPRI, 1982[10]), Boiling Water Reactor (BWR) Full-size Fine-mesh Bundle Tests (BFBT) (NEA, 2006[11]) and Pressurised Water Reactor (PWR) Subchannel and Bundle Tests (PSBT) (NEA, 2012[12]). As a preliminary activity (outside the benchmark), it is envisioned to generate the required local subchannel thermal-hydraulic parameters for all data points. The resulting subchannel output database and CHF locations (when reported in the database) will be provided to the participants.

It is expected that the participants will develop ML regression models for this task based on all previously acquired knowledge, from the base NRC CHF database in simple geometry, making use of transfer learning and applying advanced UQ methodologies. Eventually, the assessment of such an approach for reactor safety analysis, allowable operational margin, etc. will be performed and compared to current methods.

6. Timeline

The main activities related to the CHF Exercise for the Benchmark on Artificial Intelligence and Machine Learning for Scientific Computing in Nuclear Engineering are expected to be performed according to the following timeline:

CHF Benchmark introduction at Task Force meeting	December 2022
Phase 1 Draft specification and distribution	May 2023
Presentation at 2023 NEA/WPRS Annual Workshops	May 2023
Phase 1 Final specifications and distribution	September 2023
Phase 1 Online kick-off meeting	October 2023
Phase 1 Online Q&A meeting (optional)	December 2023
Phase 2 Draft specifications and distribution	May 2024
Presentation at 2024 NEA/WPRS Annual Workshops	May 2024
Phase 1 Submission	August 2024
Phase 2 Final specifications and distribution	September 2024
Phase 2 Online kick-off meeting	October 2024
Phase 1 Results draft report and online meeting	December 2024
Phase 2 (fuel bundle) Draft specifications and distribution	May 2025
Presentation at 2025 NEA/WPRS Annual Workshops	May 2025
Phase 2 Submission	August 2025
Phase 2 (fuel bundle) Final specifications and distribution	September 2025
Phase 2 (fuel bundle) Online kick-off meeting	October 2025
Presentation at 2026 NEA/WPRS Annual Workshops	May 2026
Phase 2 (fuel bundle) Submission	August 2026

References

Amazon Web Services (2023), *"Model Explainability with AWS Artificial Intelligence and Machine Learning Solutions", AWS Whitepaper*, https://docs.aws.amazon.com/whitepapers/latest/model-explainability-aws-ai-ml/interpretability-versus-explainability.htm (accessed on 24 October 2023). [9]

ASME (2022), *Verification, Validation, and Uncertainty Quantification Terminology in Computational Modeling and Simulation*, ASME VVUQ 1-2022. [19]

EPRI (1982), *Parametric Study of CHF Data, Report NP-2609*, Electric Power Research Institute (EPRI). [10]

Google (2023), *Google Machine Learning Education - Machine Learning Glossary*, https://developers.google.com/machine-learning/glossary (accessed on 1 March 2023). [17]

Groeneveld, D. (January, 2019), *Critical heat flux data used to generate the 2006 Groeneveld lookup tables*, NUREG/KM-0011. [2]

Groeneveld, D. et al. (Sept. 2007), "The 2006 CHF look-up table", *Nucl. Eng. Des.*, Vol. 237/15, pp. 1909–1922. [6]

Grosfilley, E. (2022), *Investigation of Machine Learning Regression Techniques to Predict Critical Heat Flux*, MSc thesis, Uppsala University. [4]

Grosfilley, E. et al. (2023), "Investigation of Machine Learning Regression Techniques to Predict Critical Heat Flux over a Large Parameter Space", *submitted to 20th International Topical Meeting on Nuclear Reactor Thermal Hydraulics (NURETH-20)*. [5]

Hall, D. and I. Mudawar (2000), "Critical heat flux (CHF) for water flow in tubes – II. Subcooled CHF correlations", *Int. J. Heat Mass Transf.*, Vol. 43, pp. 2605-2640. [7]

IBM (2023), *"What is machine learning?", IBM website*, https://www.ibm.com/topics/machine-learning (accessed on 1 March 2023). [15]

Kaizer, J. et al. (March 2019), *Credibility Assessment Framework for Critical Boiling Transition Models*, NUREG/KM-0013. [1]

McCarthy, J. (2004), *"What is artificial intelligence?", Stanford University*, https://cse.unl.edu/~choueiry/S09-476-876/Documents/whatisai.pdf (accessed on 1 March 2023). [14]

NEA (2023), *"Task Force on Artificial Intelligence and Machine Learning for Scientific Computing in Nuclear Engineering", NEA Website*, OECD Publishing, Paris, https://www.oecd-nea.org/jcms/pl_77779/task-force-on-artificial-intelligence-and-machine-learning-for-scientific-computing-in-nuclear-engineering (accessed on 12 September 2023). [3]

NEA (2012), "OECD/NRC Benchmark Based on NUPEC Pressurised Water Reactor Sub-channel and Bundle Tests (PSBT) - Volume I: Experimental database and final problem specifications", *NEA/NSC/DOC(2012)1, OECD Publishing, Paris*. [12]

NEA (2006), "NUPEC BWR Full-size Fine-mesh Bundle Test (BFBT) Benchmark (Volume I: Specifications)", *NEA No. 6212 (English), OECD Publishing, Paris*, https://www.oecd-nea.org/science/docs/2005/nsc-doc2005-5.pdf. [11]

NRC, U. (2023), *Artificial Intelligence Strategic Plan, Fiscal Years 2023-2027*, NUREG-2261. [13]

Preferred Networks, Inc (2023), *"Optuna - A hyperparameter optimization framework", Preferred Networks Website*, https://optuna.org/ (accessed on 5 September 2023). [8]

Roy, C. and W. Oberkampf (2011), "A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing", *Comput. Methods Appl. Mech. Engrg.*, Vol. 200, pp. 2131-2144. [18]

Trask, A. (2019), *Fundamental Concepts*, Manning Publications. [16]

Appendix A. General definitions and theory

A.1 Definitions

The benchmark involves both AI/ML and M&S terminology that could create confusion and misinterpretations. For this reason, it is important to clearly define the most important terms used within the benchmark specifications. In this appendix, definitions are provided for important AI/ML terms. The NRC has recently published a strategic plan for AI and ML that includes among others terminology definitions for various aspects of AI/ML that are relevant to this benchmark (NRC, 2023[13]). Wherever seemed appropriate, these definitions were adopted in this appendix.

Artificial Intelligence (NRC, 2023[13]) (McCarthy, 2004[14])*:* It is the science and engineering of making intelligent machines, especially intelligent computer programs, that have the ability to emulate human-like perception, cognition, planning, learning, communication, or physical action. For a given set of human-defined objectives, AI can make predictions, recommendations, or decisions influencing real or virtual environments.

Machine Learning (NRC, 2023[13]) (IBM, 2023[15]) (Trask, 2019[16])*:* It is a field of study within computer science and a branch of artificial intelligence, which focuses on developing the ability in machines to learn how to perform tasks without being explicitly programmed. Computer algorithms and data are used to gradually improve the accuracy of the performed tasks by observing underlying patterns. An illustration of the relationship between AI and ML can be found in Figure 1 of (NRC, 2023[13]).

Training (Google, 2023[17])*:* The process of determining the ML model parameters (e.g. weights) using available data. During this process the parameters are gradually updated in order to reduce the discrepancies between the predictions and the data.

Validation: The initial evaluation of the ML model accuracy using data not seen during the training process to avoid overfitting. An evaluation is performed for every set of selected hyperparameters in order to determine the optimal hyperparameters. Validation thus can be seen as the process of determining the hyperparameters for the ML model.

Testing: The final evaluation of the ML model accuracy after the hyperparameters have been selected based on the validation outcome. This evaluation is performed on data not seen neither during training nor during validation. The computed accuracy is the one associated with the ML model predictive capabilities.

Hyperparameters: Parameters that are not part of the ML model but that are part of the training process. These parameters impact the learning of the model but are not used when the model makes predictions.

Overfitting: The situation that occurs when a ML model fits the training data very closely but cannot generalise to unseen data. The validation and testing evaluations should mitigate overfitting.

Aleatoric Uncertainty (Roy and Oberkampf, 2011[18]) (ASME, 2022[19])*:* An irreducible form of uncertainty due to inherent stochastic variability. Examples of this type of uncertainty are the movement of atoms in materials and uncertainties arising in manufacturing processes.

Epistemic Uncertainty (Roy and Oberkampf, 2011[18]) (ASME, 2022[19])*:* A reducible form of uncertainty due to lack of knowledge or incomplete information. The uncertainty reduces

by gathering more information. An example of this type of uncertainty is modelling uncertainty.

Error (ASME, 2022[19])*:* The difference between a measured or calculated value and the true value or its proxy. The error can be impacted by both systematic and random effects.

A.2 Theory

Beyond terminology definitions, it is important to clearly define statistical terms used throughout this document. The mean (μ) is defined as the expected value of a random variable Y, which is assumed to be continuous with probability density function (pdf) $p_Y(y)$:

$$\mu = E[Y] = \int_{-\infty}^{\infty} y p_Y(y) dy \tag{1}$$

If a pdf is unknown and the only thing available for Y is a set of N samples/measurements y_i with $i = 1 \ldots N$, then its mean can be estimated through:

$$\hat{\mu} = \frac{\sum_{i=1}^{N} y_i}{N} \tag{2}$$

The standard deviation (σ) of a random variable is a metric of dispersion around its expected value and for the continuous random variable Y defined as:

$$\sigma = \sqrt{E[(Y-\mu)^2]} = \sqrt{\int_{-\infty}^{\infty} (y-\mu)^2 p_Y(y) dy} \tag{3}$$

As for the mean, if the only thing available for Y is a set of N samples/measurements y_i with $i = 1 \ldots N$, then the standard deviation can be estimated through:

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{\mu})^2}{N-1}} \tag{4}$$

In this benchmark, three further metrics are considered for measuring the discrepancy between predictions and measurements. The first metric is the root mean square error (RMSE) defined in Equation 5 and estimated through Equation 6. In Equation 7, the RMSE is defined in percentage. In these equations, Y^m is the measured variable with values y_i^m and Y is the predicted variable with values y_i.

$$RMSE = \sqrt{E[(Y - Y^m)^2]} \tag{5}$$

$$\widehat{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (y_i - y_i^m)^2}{N}} \tag{6}$$

$$\widehat{RMSE}_p = 100 \sqrt{\frac{\sum_{i=1}^{N} \left(\frac{y_i - y_i^m}{y_i^m}\right)^2}{N}} \tag{7}$$

The second metric is the mean absolute error (MAE) defined in Equation 8 and estimated through Equation 9. In Equation 10, the MAE is defined in percentage. MAE assigns equal weight to all discrepancies while RMSE puts more weight on larger discrepancies.

$$\mathrm{MAE} = E[|Y - Y^m|] \tag{8}$$

$$\widehat{\mathrm{MAE}} = \frac{\sum_{i=1}^{N} |y_i - y_i^m|}{N} \tag{9}$$

$$\widehat{\mathrm{MAE}}_{\mathrm{p}} = 100\frac{\sum_{i=1}^{N} \left|\frac{y_i - y_i^m}{y_i^m}\right|}{N} \tag{10}$$

Finally, the third metric is the Q^2 error (EQ^2) defined in Equation 11 and estimated through Equation 12. Both previous metrics measure discrepancies without accounting for the random variable variance. This means that small values of RMSE and MAE can be misleading when the random variable shows small variations. The EQ^2 weights the square error discrepancies by the variance and essentially measures how much of the variability of the data are actually captured by the ML model. There could be cases where the RMSE will be small and thus the numerator of EQ^2 will be small but the variance in the denominator will also be small and thus the EQ^2 metric will end up having a large value. A value of 0 indicates that the ML model has perfect predictive capabilities, while a value greater than 1 indicates that the ML model is worse than a model using always the mean value as its prediction.

$$EQ^2 = \frac{E[(Y - Y^m)^2]}{E[(Y - \mu)^2]} \tag{11}$$

$$EQ^2 = \frac{\sum_{i=1}^{N}(y_i - y_i^m)^2}{\sum_{i=1}^{N}(y_i - \hat{\mu})^2} \tag{12}$$

Appendix B. Files

All required input files are stored in the NEA GitLab repository of the NEA NSC/WPRS/EGMUP Task Force on AI and Machine Learning: in directory: https://git.oecd-nea.org/science/wprs/egmup-task-force-on-ai-and-machine-learning/phase1-chf-exercises/-/tree/main/data/.

Access to the NEA GitLab repository is restricted to registered benchmark participants.

Table B.1. List of CHF benchmark files

Directory name	File name	Description
inputs	chf_public.csv	NRC CHF database
	chf_blind.csv	Blind test data set
	model_summary.xlsx	Model and evaluation summary template
	Slice_01.csv	Slice data set with varying D (1)
	Slice_02.csv	Slice data set with varying D (2)
	Slice_03.csv	Slice data set with varying L (1)
	Slice_04.csv	Slice data set with varying L (2)
	Slice_05.csv	Slice data set with varying P (1)
	Slice_06.csv	Slice data set with varying P (2)
	Slice_07.csv	Slice data set with varying G (1)
	Slice_08.csv	Slice data set with varying G (2)
	Slice_09.csv	Slice data set with varying X (1)
	Slice_10.csv	Slice data set with varying X (2)
results/LUT	LUT2006.xls	Tabulated LUT
	chf_public_LUT.csv	LUT results for the NRC CHF database
	Slice_01_LUT.csv	LUT results for the slice data set with varying D (1)
	Slice_02_LUT.csv	LUT results for the slice data set with varying D (2)
	Slice_03_LUT.csv	LUT results for the slice data set with varying L (1)
	Slice_04_LUT.csv	LUT results for the slice data set with varying L (2)
	Slice_05_LUT.csv	LUT results for the slice data set with varying P (1)
	Slice_06_LUT.csv	LUT results for the slice data set with varying P (2)
	Slice_07_LUT.csv	LUT results for the slice data set with varying G (1)
	Slice_08_LUT.csv	LUT results for the slice data set with varying G (2)
	Slice_09_LUT.csv	LUT results for the slice data set with varying X (1)
	Slice_10_LUT.csv	LUT results for the slice data set with varying X (2)
results/NN	chf_public_NN.csv	Example of NN results for the NRC CHF database
	model_summary_NN.xlsx	Example of NN model and evaluation summary
	Slice_01_NN.csv	NN results for the slice data set with varying D (1)
	Slice_02_NN.csv	NN results for the slice data set with varying D (2)
	Slice_03_NN.csv	NN results for the slice data set with varying L (1)
	Slice_04_NN.csv	NN results for the slice data set with varying L (2)
	Slice_05_NN.csv	NN results for the slice data set with varying P (1)
	Slice_06_NN.csv	NN results for the slice data set with varying P (2)
	Slice_07_NN.csv	NN results for the slice data set with varying G (1)
	Slice_08_NN.csv	NN results for the slice data set with varying G (2)
	Slice_09_NN.csv	NN results for the slice data set with varying X (1)
	Slice_10_NN.csv	NN results for the slice data set with varying X (2)

Appendix C. CHF Lookup (LUT) table results

The CHF lookup table (LUT) (Groeneveld et al., Sept. 2007[6]) is a simple data-driven model to predict CHF in vertical uniformly heated tubes over a wide range of conditions. For practical use of a tabulated model, the number of input data entries is limited to three. Since they are generally seen as the most influential parameters, the LUT utilises the pressure, *P*, mass flux, *G*, and local equilibrium quality, *X*, as input parameters within ranges shown in the following table.

Table C.1. Parameter ranges covered by the CHF LUT

PARAMETER	MINIMUM	STEP-SIZE	MAXIMUM
P [kPa]	100	200-3000	21000
G [kg/m^2/s]	0	200-500	8000
X [-]	-0.5	0.05-0.1	1.0

For any given triplets (*P*, *G*, *X*), linear interpolation within the LUT is used to predict CHF for a reference tube diameter of 8 mm. The following correction is then performed to adjust the predicted CHF to the actual tube diameter (Groeneveld et al., Sept. 2007[6]):

$$CHF_D = CHF_{8\,mm} \cdot \left(\frac{D}{8\,mm}\right)^{-0.5}$$

where *D* can range within values covered by the database (that is, 2 to 16 mm).

The CHF database used to develop the LUT is roughly the same as the NRC CHF database, with the exception of some additional (but limited) proprietary CHF data. The LUT is reported to have a root mean squared error of 38.92% when predicting CHF using constant local conditions (Groeneveld et al., Sept. 2007[6]).

The results of the CHF LUT for the relevant benchmark activities are documented in this appendix. All data are provided in the formats required for the benchmark and provided in the following directory on NEA GitLab[2]:

https://git.oecd-nea.org/science/wprs/egmup-task-force-on-ai-and-machine-learning/phase1-chf-exercises/-/tree/main/data/results/LUT.

This can be used both as reference results (that is, the developed ML models are expected to outperform the LUT) and as example of data and information to be provided.

In the case where the participants would like to perform their own evaluation using the CHF LUT, the 8 mm CHF tables are provided in the file `LUT2006.xls`.

The CHF predictions for the 24 579 data points of the NRC CHF database are provided in the file `CHF_public_LUT.csv`. An example (when open in Excel) for the first 5 data points is shown in Table C.2. The corresponding model performance metrics calculated for the entire NRC CHF database (24 579 data points) are provided below in Table C.3. The model prediction plots are provided in Figure C.1 (measured CHF vs predicted CHF), Figure C.2 (P/M histogram) and Figure C.3 (P/M scatter plots vs selected independent parameters).

[2] Access to NEA GitLab is restricted to registered benchmark participants.

The results from the slice analysis are shown in the following pages. An example of LUT CHF prediction for slice 1 (varying diameter) can be found in Table C.4. The plots for all considered slices are shown in Figures C.4 to C.8. The experimental data at similar conditions found within the NRC CHF database are represented by circle markers.

Table C.2. Example of CHF LUT results

Number	Reference ID	Tube Diameter	Heated Length	Pressure	Mass Flux	Outlet Quality	Inlet Subcooling	Inlet Temperature	CHF	CHF LUT
-	-	m	m	kPa	kg/m^2/s	-	kJ/kg	C	kW/m^2	kW/m^2
1	1	0.004	0.396	100	77.5	0.84	317	23.94	442	469.0946386
2	1	0.004	0.396	100	142.7	0.79	317	23.94	757	687.0899317
3	1	0.004	0.396	100	203.9	0.7	317	23.94	978	903.6294333
4	1	0.004	0.396	100	271.8	0.73	317	23.94	1325	873.6220843
5	1	0.004	0.396	100	421.3	0.62	317	23.94	1798	1372.801147

Table C.3. CHF LUT prediction performances

All	Size of data set		24579 samples
	Evaluation	Mean P/M	1.032
		Std P/M	0.362
		RMSPE [%]	19.8
		MAPE [%]	36.300
		Q^2 error	0.063

Figure C.1. Measured vs LUT predicted CHF

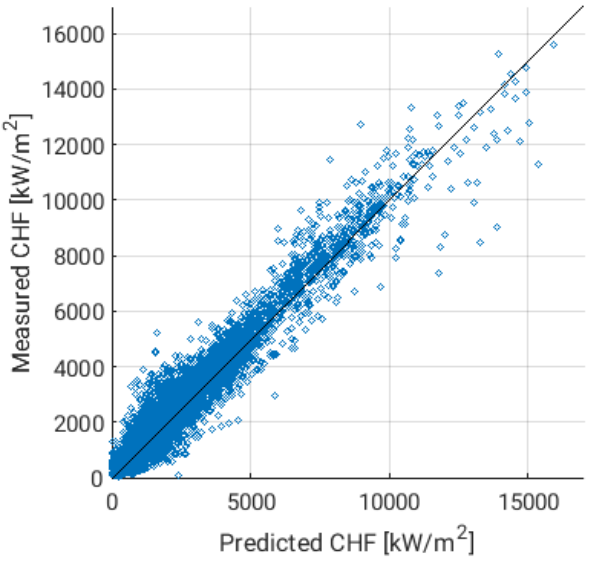


Figure C.2. LUT Predicted over measured CHF histogram

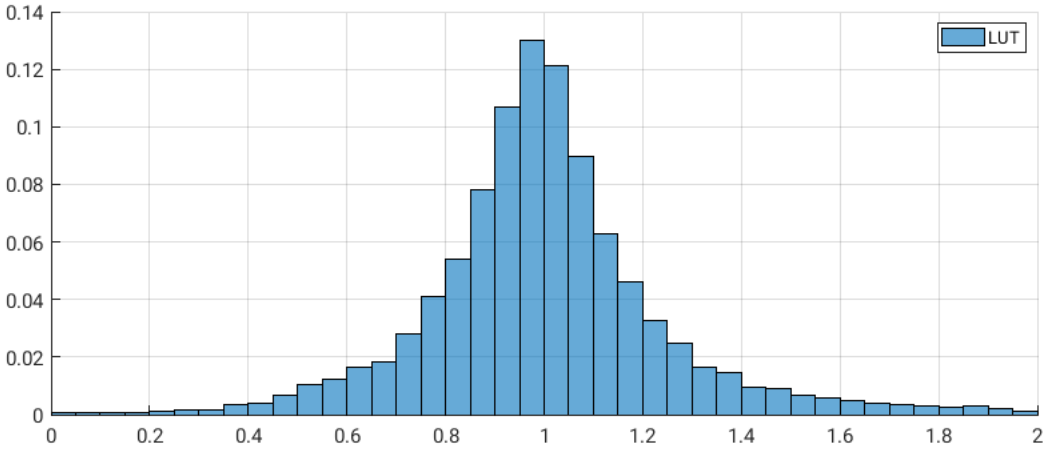


Figure C.3. LUT Predicted over Measured CHF scatter plots, vs selected independent parameters

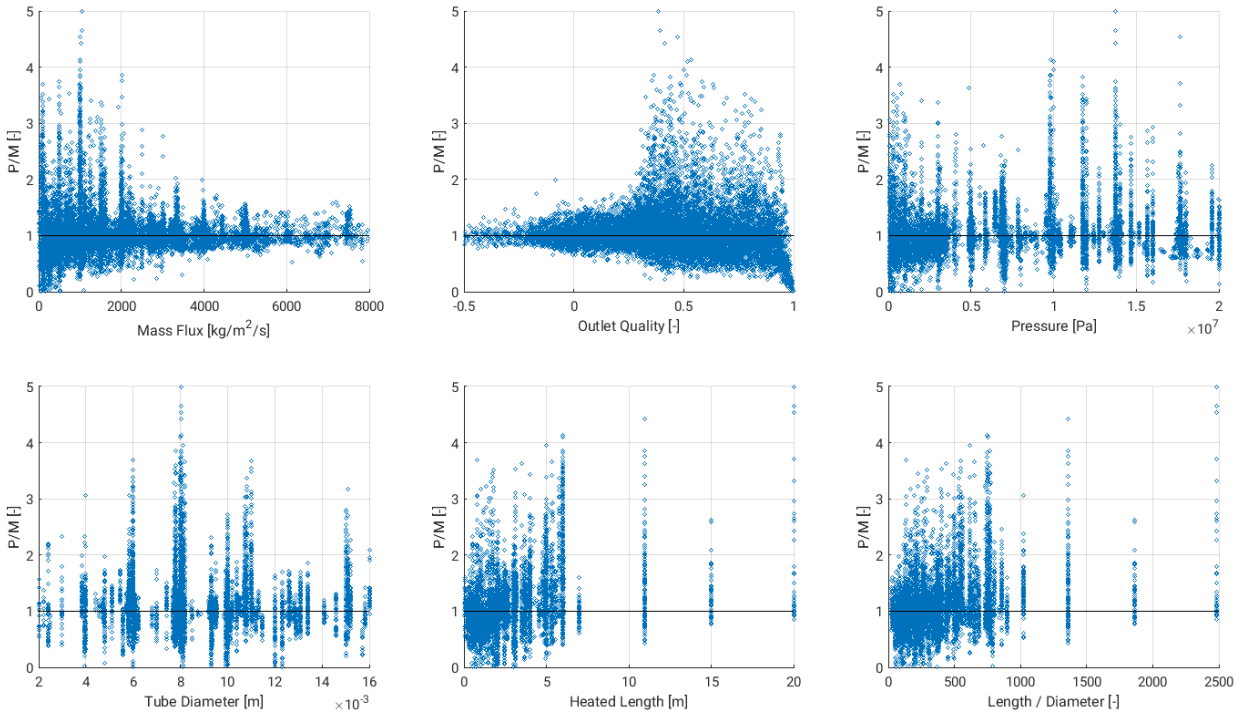


Table C.4. CHF LUT prediction results for Slice 1

Tube Diameter	Heated Length	Pressure	Mass Flux	Outlet Quality	CHF LUT
m	m	Pa	kg/m^2/s	-	W/m^2
0.002	6	14701250	998.5	0.390875	1258364.212
0.003	6	14701250	998.5	0.390875	1027450.076
0.004	6	14701250	998.5	0.390875	889797.8672
0.005	6	14701250	998.5	0.390875	795859.4069
0.006	6	14701250	998.5	0.390875	726516.9163
0.007	6	14701250	998.5	0.390875	672623.9639
0.008	6	14701250	998.5	0.390875	629182.1058
0.009	6	14701250	998.5	0.390875	593198.5781
0.01	6	14701250	998.5	0.390875	562757.5835
0.011	6	14701250	998.5	0.390875	536568.3027
0.012	6	14701250	998.5	0.390875	513725.0381
0.013	6	14701250	998.5	0.390875	493571.0515
0.014	6	14701250	998.5	0.390875	475616.9661
0.015	6	14701250	998.5	0.390875	459489.6428
0.016	6	14701250	998.5	0.390875	444898.9336

Figure C.4. Variations of predicted LUT CHF vs diameter (D, in [m]) for slices 1 and 2, and corresponding NRC CHF data points

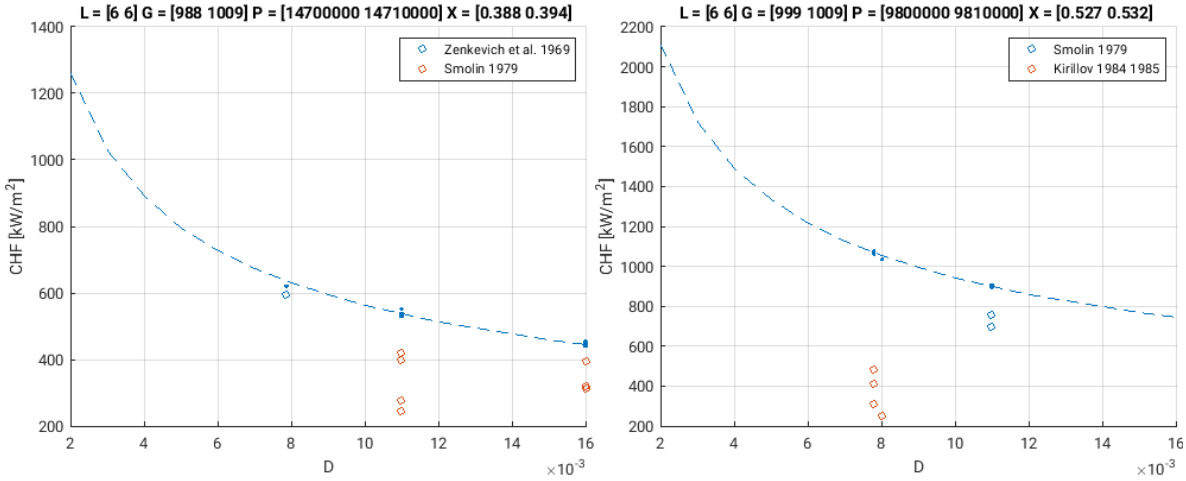


Figure C.5. Variations of predicted LUT CHF vs heated length (*L*, in [m]) range for slices 3 and 4, and corresponding NRC CHF data points

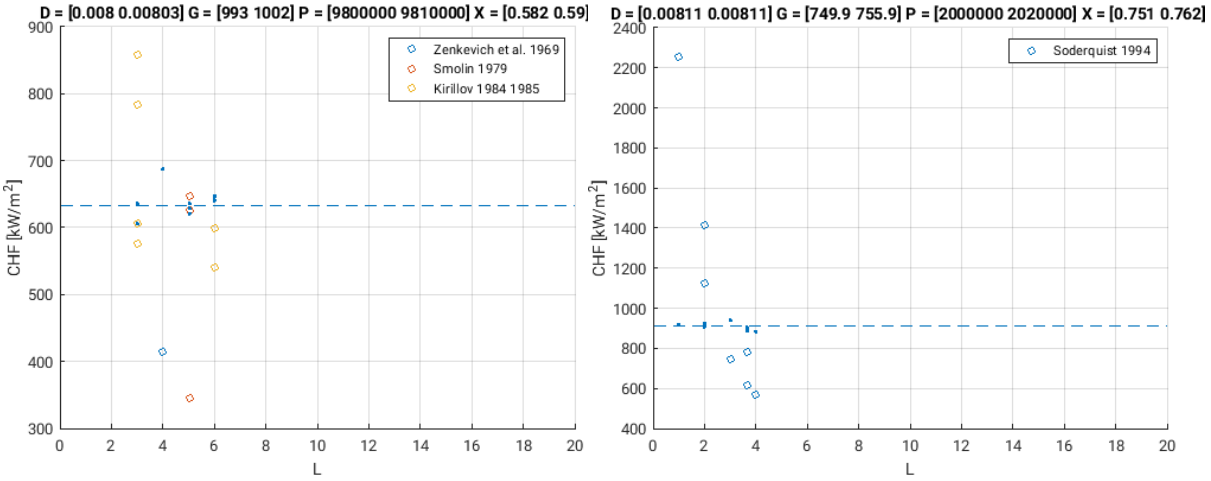


Figure C.6. Variations of predicted LUT CHF vs pressure (*P*, in [Pa]) for slices 5 and 6, and corresponding NRC CHF data points

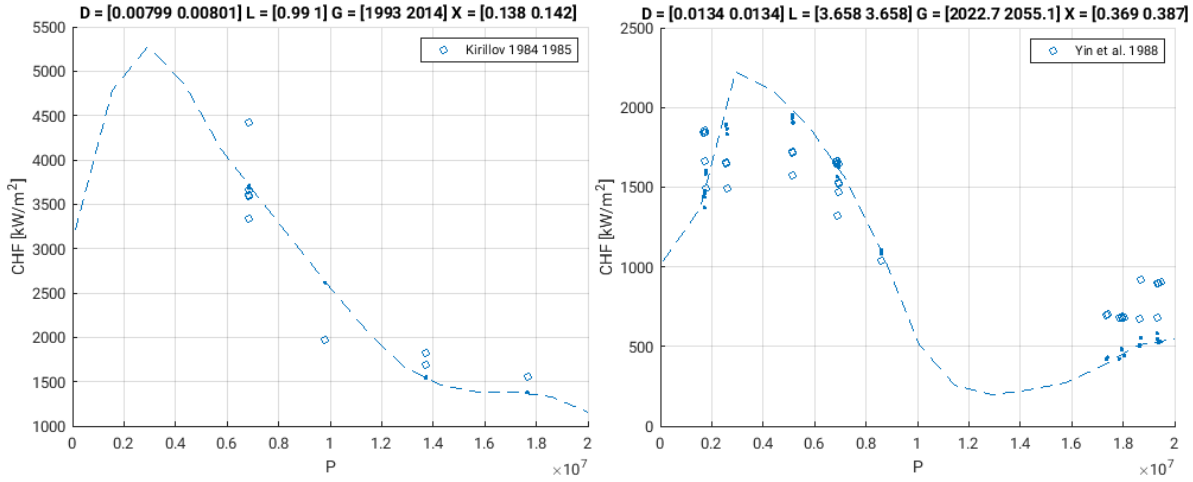


Figure C.7. Variations of predicted LUT CHF vs mass flux (*G*, in [kg/m^2/s]) for slices 7 and 8, and corresponding NRC CHF data points

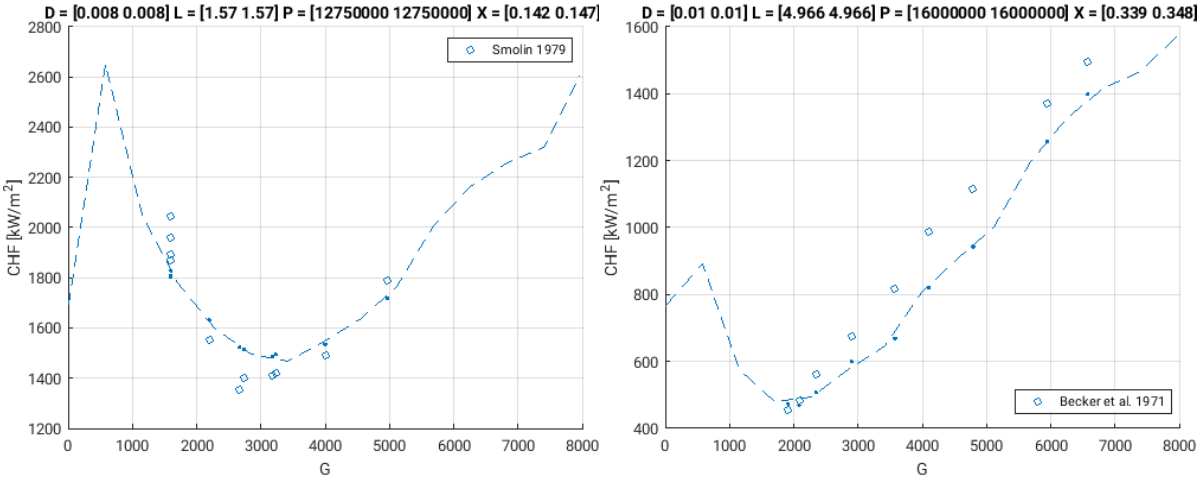
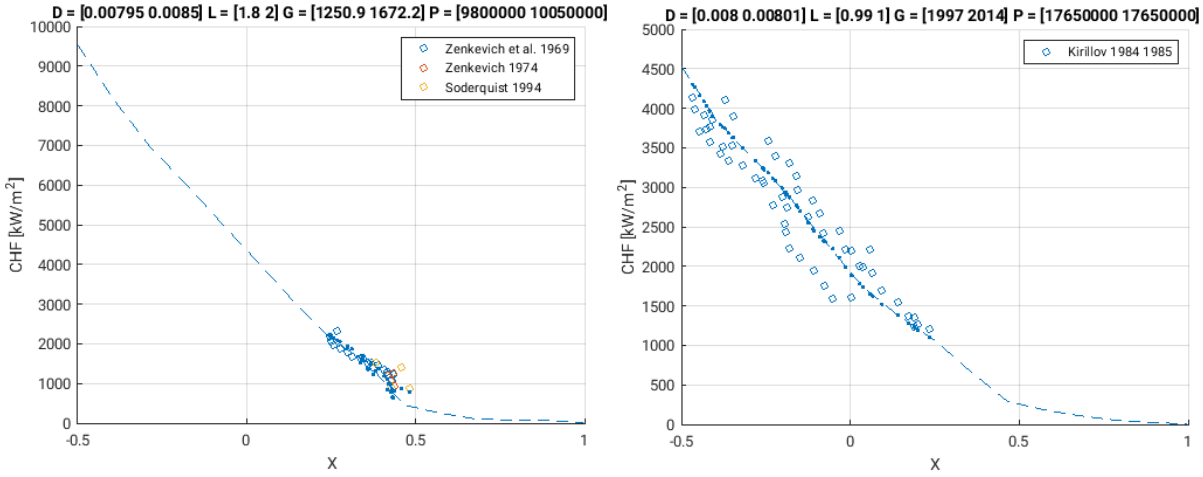


Figure C.8. Variations of predicted LUT CHF vs quality (*X*) for slices 9 and 10 and corresponding NRC CHF data points



Appendix D. Example of neural network (NN) results

A neural network (NN) was developed in (Grosfilley, 2022[4]) and (Grosfilley et al., 2023[5]), selected among various tested ML algorithms, based on the CHF NRC database. The requested benchmark outputs have been generated for this model and documented in this appendix. The performances of the model have also been compared to the reference CHF LUT predictions.

Note that this predictive CHF model uses the heated length, *L*, as input. Unless justified, direct use of this parameter is, however, not recommended (see discussion in Section 3.1.1).

The ML model development tools and architecture, hyperparameters, training/optimisation methods and calculated model performance metrics are documented in Table D.1.

The NN was trained over 1 150 epochs where the loss (MSLE) decreased as shown in Figure D.1. Note that the loss for the test database was only calculated once, after completion of the training.

The results obtained with the NN is provided in the following directory on NEA GitLab[3]:

https://git.oecd-nea.org/science/wprs/egmup-task-force-on-ai-and-machine-learning/phase1-chf-exercises/-/tree/main/data/results/NN.

The CHF predictions for the 24 579 data points of the NRC CHF database are provided in the file `CHF_public_NN.csv`.

An example (when open in Excel) for the first 5 data points is shown in Table D.2.

The model prediction plots are provided in Figure D.2 (measured CHF vs predicted CHF), Figure D.3 (P/M histogram) and Figure D.4 (P/M scatter plots vs selected independent parameters). For all plots, the prediction of both the LUT (also shown in Appendix C) and the NN are shown.

An example of NN CHF prediction for slice 1 (varying diameter) can be found in Table D.3. The plots for all considered slices are shown in Figures D.5 to D.9. The experimental data at similar conditions found within the NRC CHF database are represented by circle markers. For all plots, the results from the LUT are also represented for comparison.

[3] Access to NEA GitLab is restricted to registered benchmark participants.

Table D.1. Example of CHF NN architecture, hyperparameters and prediction performances

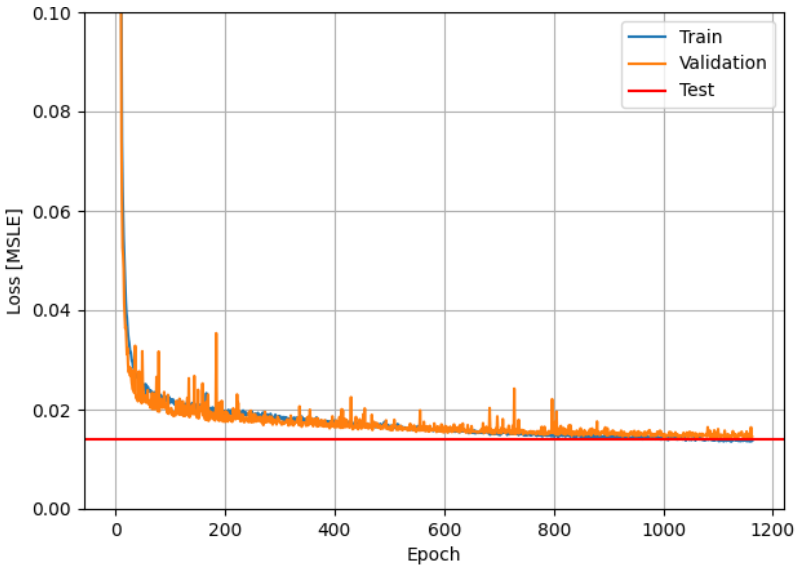
General		Algorithm	Neural network
		Implementation platform	TensorFlow
		Hyperparameter optimization	Optuna
		Hyperparameter opt. target	MSLE and size
		Inputs	P, G, D, L, X
		Output	CHF
		Layer architecture	5, 61, 51, 28, 39, 26, 21, 20, 14, 1
		Activation function	ReLu
Training	Hyperparameters	Data scaling	StandardScaler
		Batch size	16
		Training steps per epoch	1230
		Optimizer	Adam
		Regularization	Dropout rate of 0.01 on first two layers
		Learning rate	0.001
		Learning rate decay factor	0.96 every 32 epoch
		Loss function	MSLE
		Validation	5-fold cross-validation
		Early stopping	No decrease in validation loss for 50 epochs in a row
		Number of epochs	1150
	Evaluation	Size of data set	Random 80% of 24579 samples
		Mean P/M	1.009
		Std P/M	0.113
		RMSPE [%]	12.5
		MAPE [%]	
		Q^2 error	
Test	Evaluation	Size of data set	Remaining 20% of 24579 samples
		Mean P/M	1.013
		Std P/M	0.123
		RMSPE [%]	13.4
		MAPE [%]	
		Q^2 error	
All	Evaluation	Size of data set	24579 samples
		Mean P/M	1.010
		Std P/M	0.115
		RMSPE [%]	12.6
		MAPE [%]	8.0
		Q^2 error	0.022

Source: (Grosfilley, 2022[4]) (Grosfilley et al., 2023[5])

Table D.2. Example of CHF NN results

Number	Reference ID	Tube Diameter	Heated Length	Pressure	Mass Flux	Outlet Quality	Inlet Subcooling	Inlet Temperature	CHF	CHF NN
-	-	m	m	kPa	kg/m^2/s	-	kJ/kg	C	kW/m^2	kW/m^2
1	1	0.004	0.396	100	77.5	0.84	317	23.94	442	485.8004
2	1	0.004	0.396	100	142.7	0.79	317	23.94	757	838.8955
3	1	0.004	0.396	100	203.9	0.7	317	23.94	978	1001.399
4	1	0.004	0.396	100	271.8	0.73	317	23.94	1325	1414.337
5	1	0.004	0.396	100	421.3	0.62	317	23.94	1798	1867.432

Figure D.1. NN loss curve



Source: reproduced from (Grosfilley et al., 2023[5]).

Figure D.2. Measured vs LUT (blue) and NN (green) predicted CHF

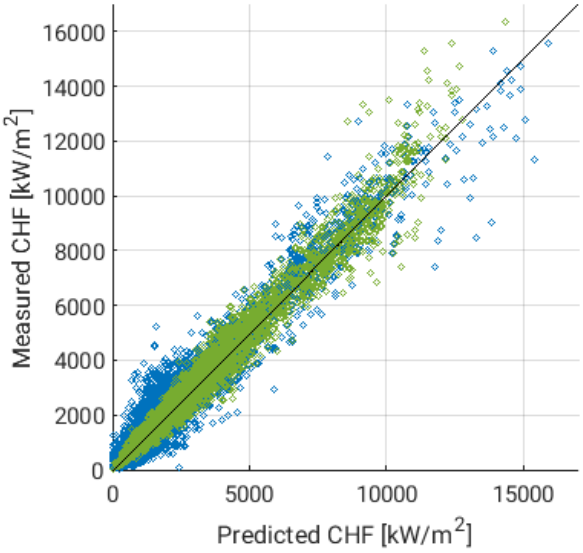


Figure D.3. LUT (blue) and NN (green) Predicted over Measured CHF histogram

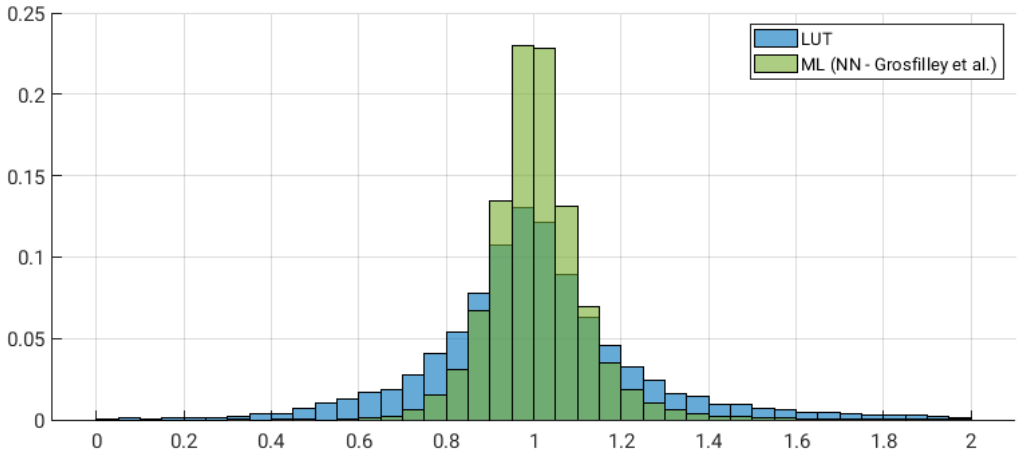


Figure D.4. LUT (blue) and NN (green) Predicted over Measured CHF scatter plots, vs selected independent parameters

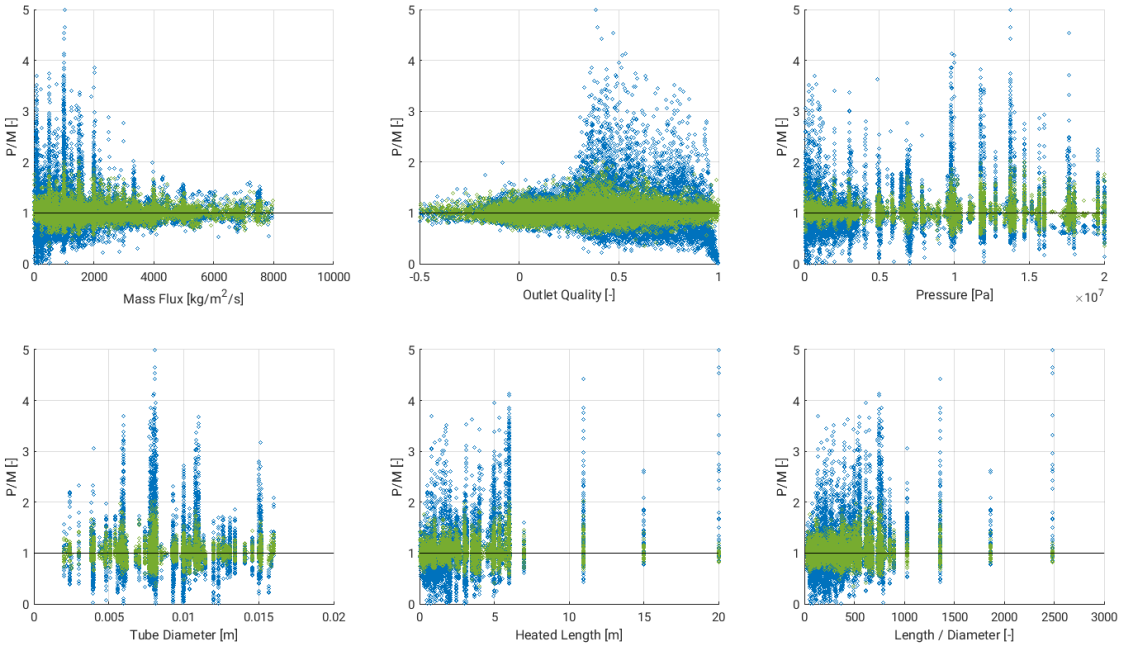


Table D.3. NN CHF prediction results for Slice 1

Tube Diameter	Heated Length	Pressure	Mass Flux	Outlet Quality	CHF LUT
m	m	Pa	kg/m^2/s	-	W/m^2
0.002	6	14701250	998.5	0.390875	657659.8
0.003	6	14701250	998.5	0.390875	550817
0.004	6	14701250	998.5	0.390875	343829.4
0.005	6	14701250	998.5	0.390875	305005.5
0.006	6	14701250	998.5	0.390875	357454.7
0.007	6	14701250	998.5	0.390875	384713
0.008	6	14701250	998.5	0.390875	450119.7
0.009	6	14701250	998.5	0.390875	442271.2
0.01	6	14701250	998.5	0.390875	449112.6
0.011	6	14701250	998.5	0.390875	377457.9
0.012	6	14701250	998.5	0.390875	397470.5
0.013	6	14701250	998.5	0.390875	425027.6
0.014	6	14701250	998.5	0.390875	475281.9
0.015	6	14701250	998.5	0.390875	483809
0.016	6	14701250	998.5	0.390875	438802.7

Figure D.5. Variations of predicted LUT (blue) and NN (green) CHF vs diameter (D, in [m]) for slices 1 and 2, and corresponding NRC CHF data points

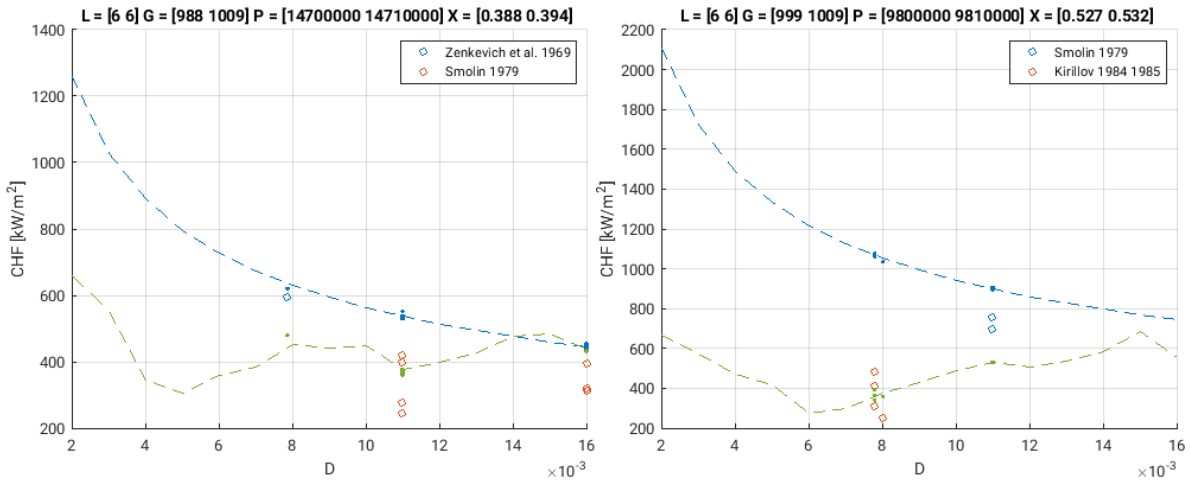


Figure D.6. Variations of predicted LUT (blue) and NN (green) CHF vs heated length (*L*, in [m]) for slices 3 and 4, and corresponding NRC CHF data points

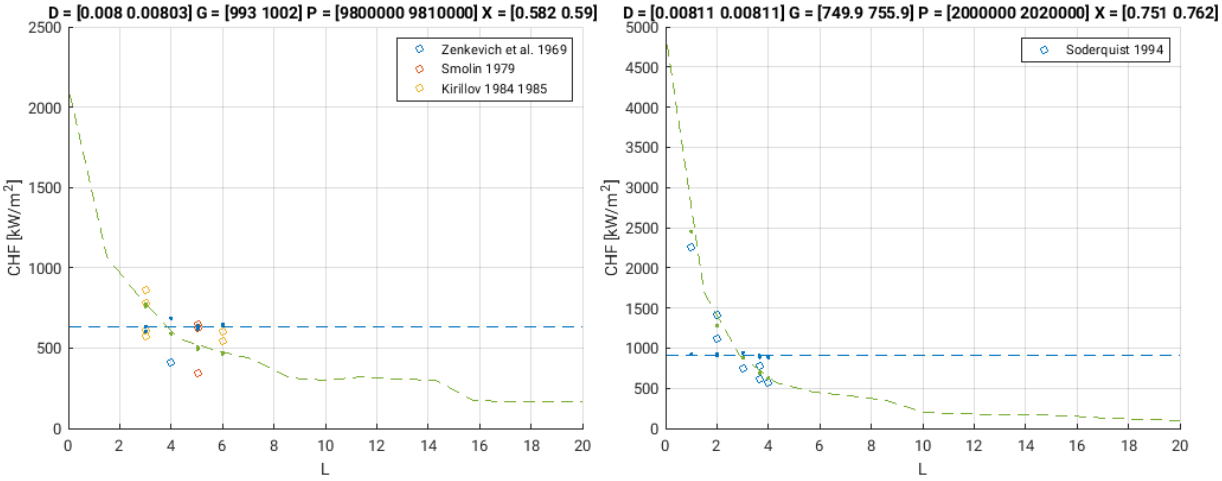


Figure D.7. Variations of predicted LUT (blue) and NN (green) CHF vs pressure (*P*, in [Pa]) for slices 5 and 6, and corresponding NRC CHF data points

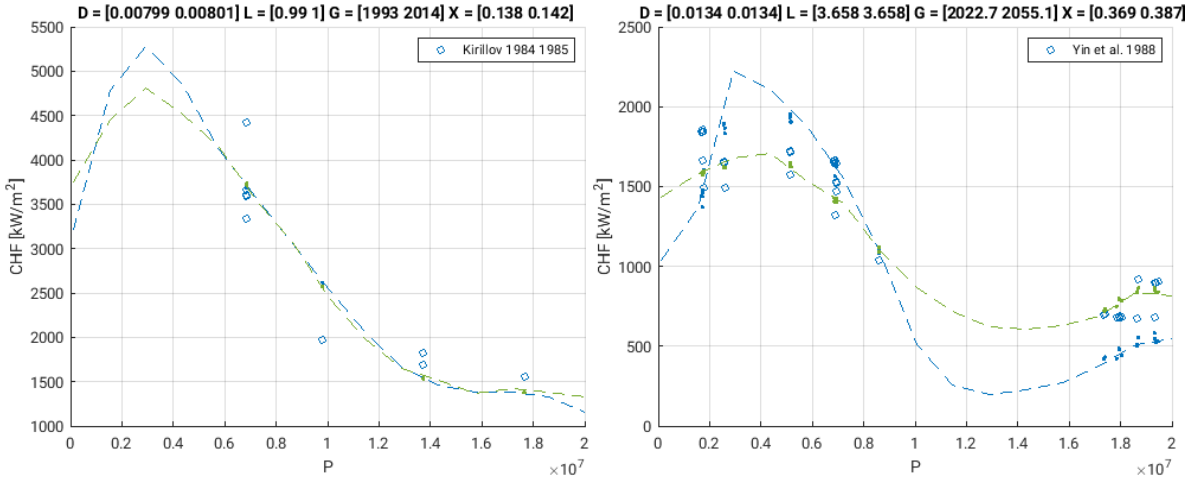


Figure D.8. Variations of predicted LUT (blue) and NN (green) CHF vs mass flux (*G*, in [kg/m^2/s]) for slices 7 and 8, and corresponding NRC CHF data points

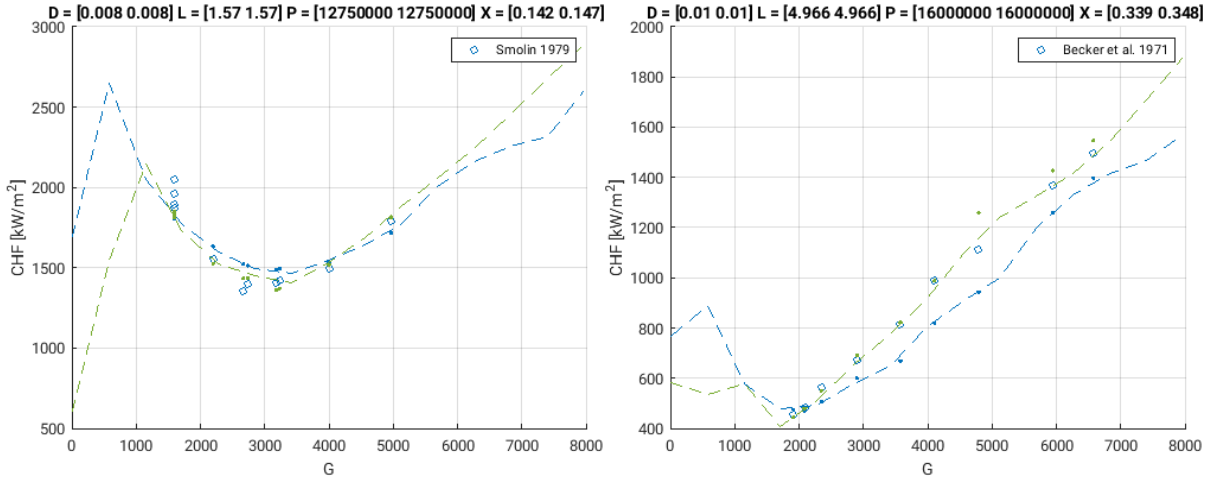


Figure D.9. Variations of predicted LUT (blue) and NN (green) CHF vs quality (*X*) for slices 9 and 10, and corresponding NRC CHF data points