Utilizing Machine Learning for Phenomena-based Synthesis of Intensified Process Flowsheets: Supplementary Material

Omar Alqusair[ab]*, Jie Li[a]

[a] Centre for Process Integration, Department of Chemical Engineering, The University of Manchester, Manchester, M13 9PL, UK
[b] Department of Chemical Engineering, College of Engineering, King Saud University, P.O. Box 800, Riyadh 11421, Saudi Arabia
* Corresponding Author: omar.alqusair@manchester.ac.uk.

Heuristic and Sampling Logic Rules

Table S.1: Heuristic rules used in generating the initial dataset.

Type	#	Rule Description
Grammar	G1	Input and desired output tokens must appear at least once in the string.
	G2	A string must end with either an output token or an recycle outlet token.
	G3	Input, output, and phenomena tokens use '(' to denote the start of the token, and ')' for its end.
	G4	The next token after a mixer token is either a mixer branch start or a recycle inlet.
	G5	Mixer branches use the '<&|' to denote the branch's start, and the '&|' for the branch's end.
	G6	The first token inside a mixing branch must be an input token or a recycle input token.
	G7	The next token after a separation token is either a splitting branch start or a recycle outlet.
	G8	Splitting branches use '[' to denote the branch's start, and ']' for the branch's end.
	G9	The last token inside a splitting branch must either be an output or recycle outlet.
	G10	Recycle streams use '<#' to denote a recycle inlet, and '#' for its outlet, where # is an integer.
	G11	A recycle outlet is only allowed if its corresponding recycle inlet appears in the string.
	G12	Recycle inlet and outlet pairs can appear at most once.
	G13	The last token in a string can either be an output, or a recycle outlet
Thermodynamic	T1	If an output token is not available as an input, then at least one reaction token must exist..
	T2	Non-mixing reactor tokens require at least 2 inputs to appear before it in the string
	T3	Input tokens that are involved in chemical reactions must appear before the first reaction token.
	T4	If a non-separation reaction token appears in the string, at least one separation token must appear in the string afterwards.
	T5	One-phenomenon tokens of the same subtype cannot appear consecutively in the string.
	T6	A recycle outlet is only allowed after the second input token appear in the string.

Table S.2: Sampling logic rules used in generating the initial dataset.

#	Rule Description
S1	The maximum number of tokens in a string is a random value between [#inputs+#outputs+1, n]
S2	The first token in a string is a random unused input token
S3	Before each token placement, refresh the list of [eligible] tokens, to filter out ineligible tokens
S4	The tokens inside the [eligible] list are further partitioned into mixing/splitting/other buckets
S5	For each position in the string, a token is picked from a non-empty bucket.
S6	If a mixing or separation token is placed, the [eligible] list is overridden to enforce branch start.
S7	If a branch start token is added, randomly pick a nest length between 1 and 3, and fill-in the nest.
S8	Recycle inlet and outlet pairs are unique and cannot be duplicated in a string.
S19	Recycle outlets can only appear after an input token has been placed.
S10	The targeted number of mixer branches is equal to the number of input tokens – 1.
S11	The targeted number of splitter branches is equal to the number of output tokens – 1.
S12	The minimum number of tokens inside a branch is 1, and the maximum number is 4.
S13	For each position in a string, the next token to be sampled is taken from a list of eligible tokens.
S14	All open branches must be terminated using an appropriate branch end token.
S15	If the current string length plus pending outlets would exceed the maximum number of tokens, terminate and flag as infeasible.
S16	A string is flagged feasible only if the string follows the heuristic rules and meets the branch and recycle targets.

Hyperparameter Grid Search

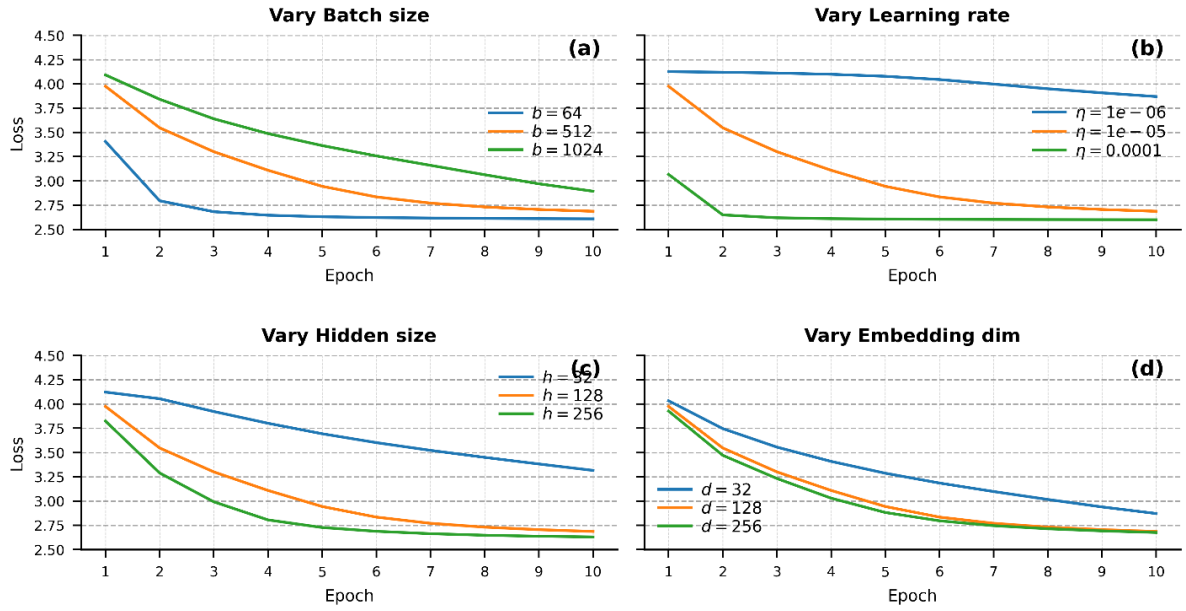

Figure S.1. Training cross-entropy loss curves by varying: (a) batch size, (b) learning rate, (c) hidden size, and (d) embedding dimension.

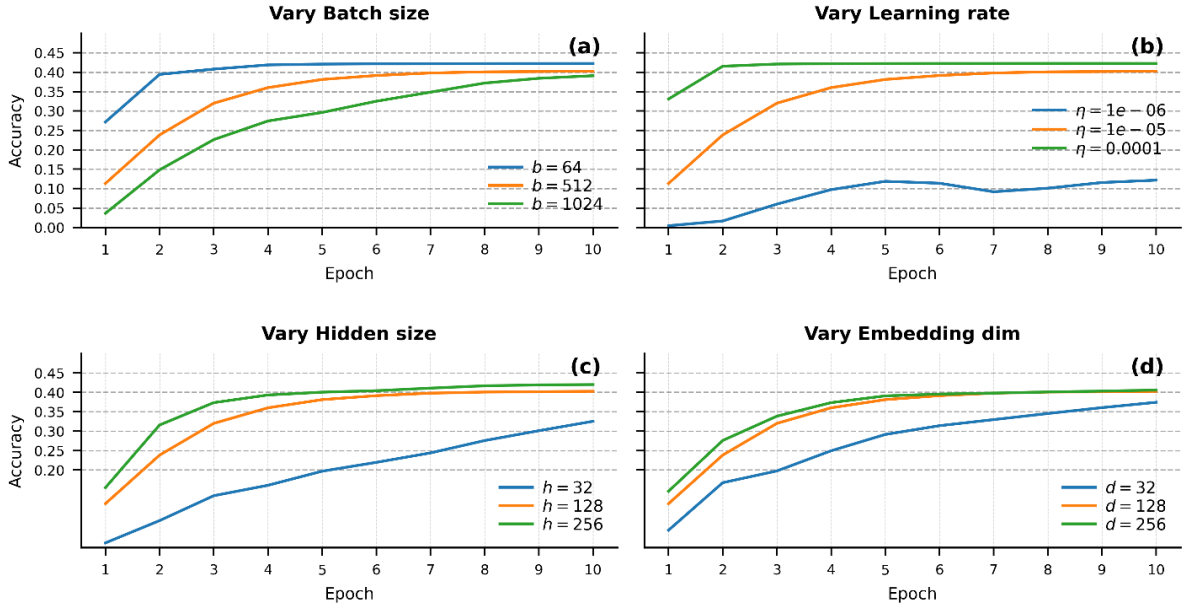

Figure S.2. Training accuracy curves by varying: (a) batch size, (b) learning rate, (c) hidden size, and (d) embedding dimension.