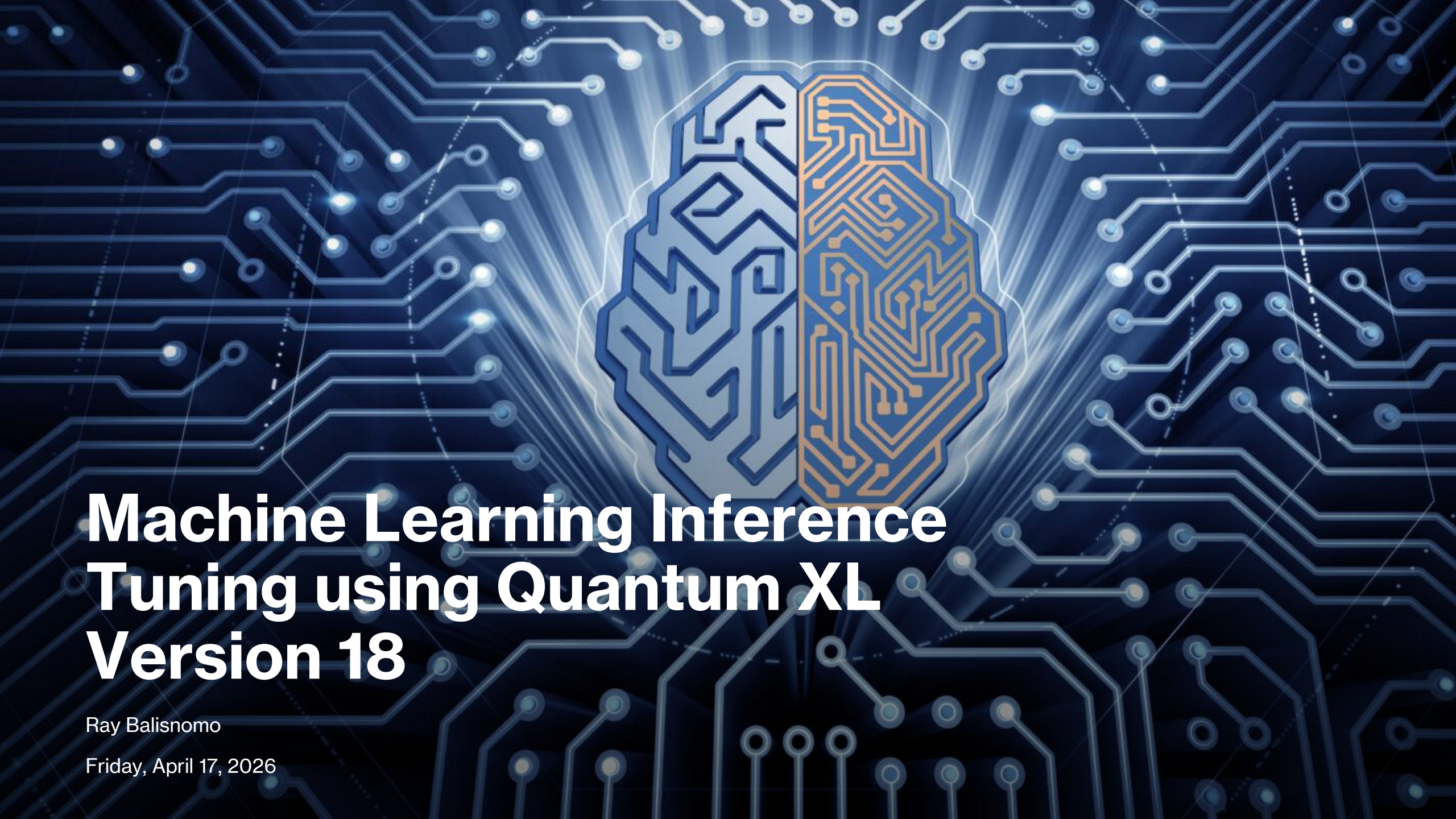


Machine Learning Inference Tuning using Quantum XL Version 18

Ray Balisnomo

Friday, April 17, 2026

Introduction

- Machine Learning (ML) inference tuning is the systematic optimization of how a trained model executes in production to achieve predictable, robust performance while respecting accuracy, resource, and reliability constraints.

- ML inference tuning is the practice of deliberately adjusting how a trained machine learning model is executed (not how it is trained) to achieve the best trade off between performance, accuracy, resource usage, and robustness in a real deployment.

- Think of it as engineering optimization, not data science.

- Inferencing tuning asks: “What settings work **consistently**, not just once?”

ML Inference Tuning vs. “Benchmark Chasing”

BENCHMARK CHASING

- Maximize one metric
- Best-case scenario
- Fixed workload
- Hero numbers (works just once)

INFERENCE TUNING

- Balance multiple objectives
- Robust across conditions
- Workload variability
- Predictable performance (works consistently)

Why engineers increasingly use DOE for inference tuning

1) Because inference behavior is:
   - **nonlinear**
   - **hardware dependent**
   - **interaction heavy**
2) Response Surface Design allows you to:
   - **quantify trade offs**
   - **predict unseen settings**
   - **find robust operating regions**
   - **defend decisions with data**
3) It turns ML inference from **trial and error** into **engineering discipline**.

Problem Context

You are deploying an ML model (CNN / transformer / classical ML) in a firmware or software runtime where you must balance:

- **inference latency**
- **throughput**
- **accuracy**
- **CPU / accelerator utilization**
- **power consumption**

These tradeoffs are nonlinear and highly interactive, so OFAT tuning (one factor at a time) fails.

Factors and Engineering Ranges

- The ranges below are chosen to reflect what engineers can actually change without redesigning the model.
- Control factors: Response Surface Design (RSD) does **not** allow categorical factors.

Factor	Symbol	Type	Low	Center	High
Batch Size	**A**	**Numeric**	**1**	**8**	**16**
Quantization	**B**	**Ordinal**	**INT4**	**INT8**	**FP16**
Precision Mode	**C**	**Ordinal**	**FP16**	**TF32**	**FP32**

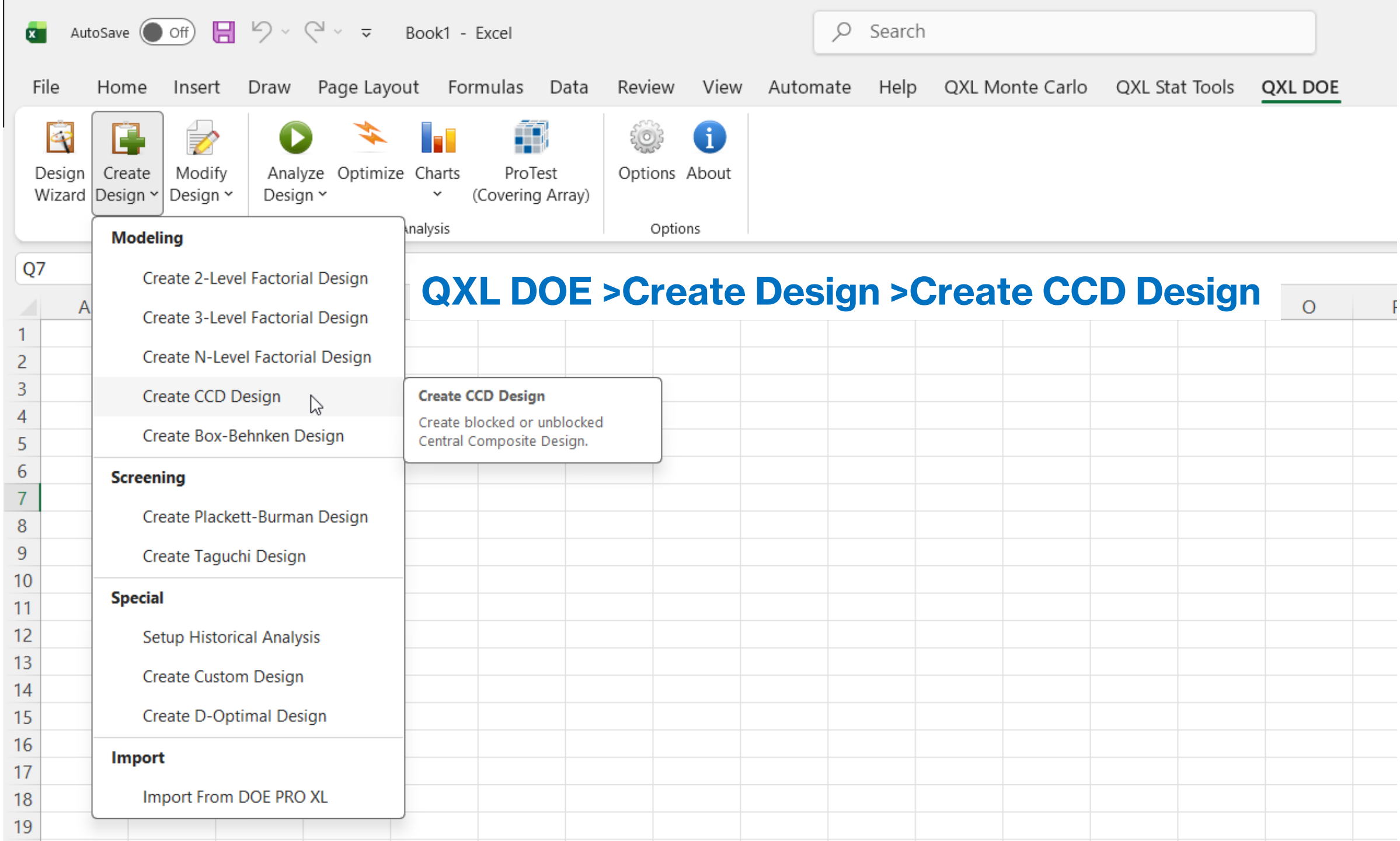
Encoding Ordinal Factors

- Although quantization and precision are categorical, they are **ordered by numeric fidelity** and can be treated as **ordinal numeric factors** for RSD.
- This is common practice when strategies represent increasing strength or fidelity.

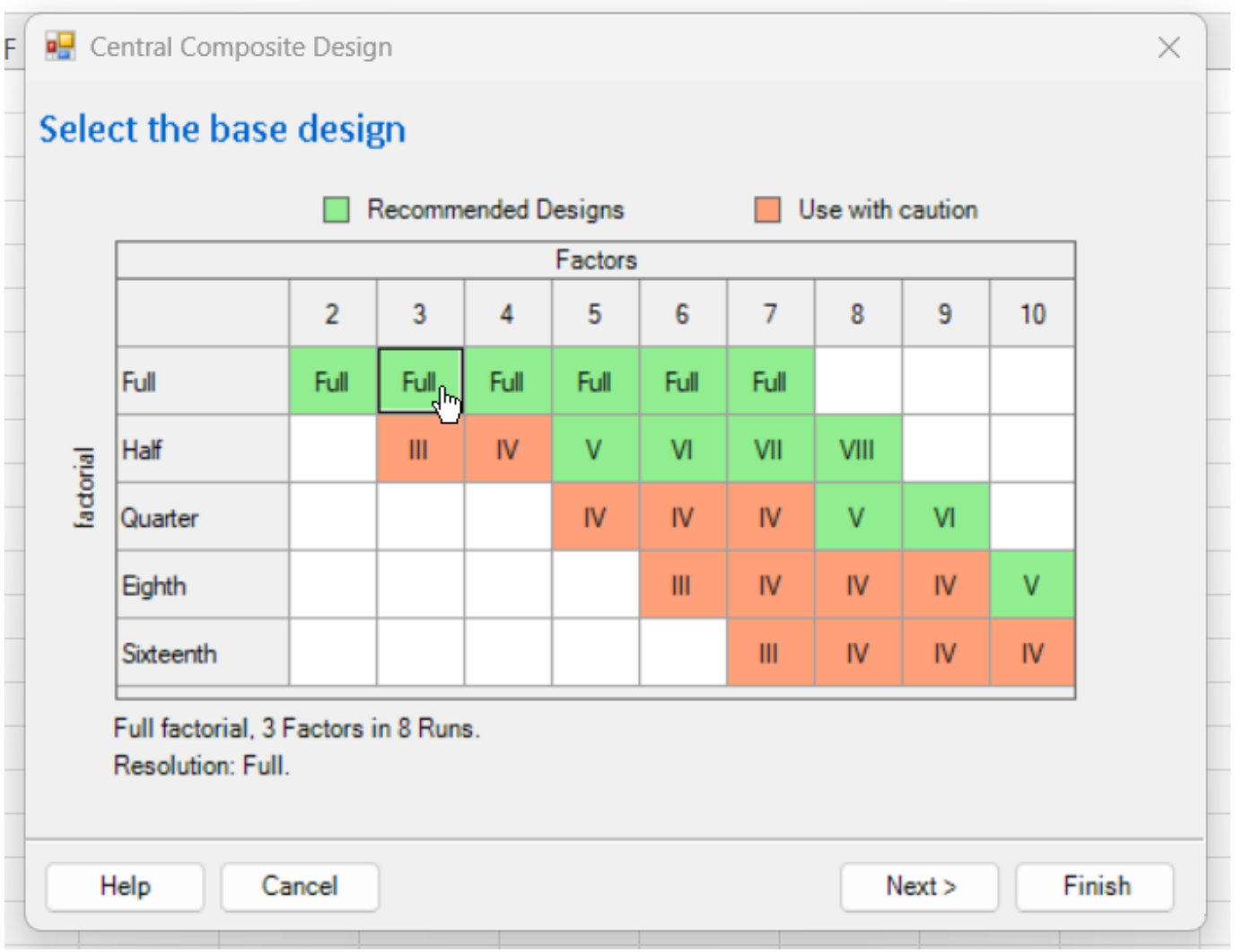
Level	Quantization	Precision Mode
–1	**INT4**	**FP16**
0	**INT8**	**TF32**
+1	**FP16**	**FP32**

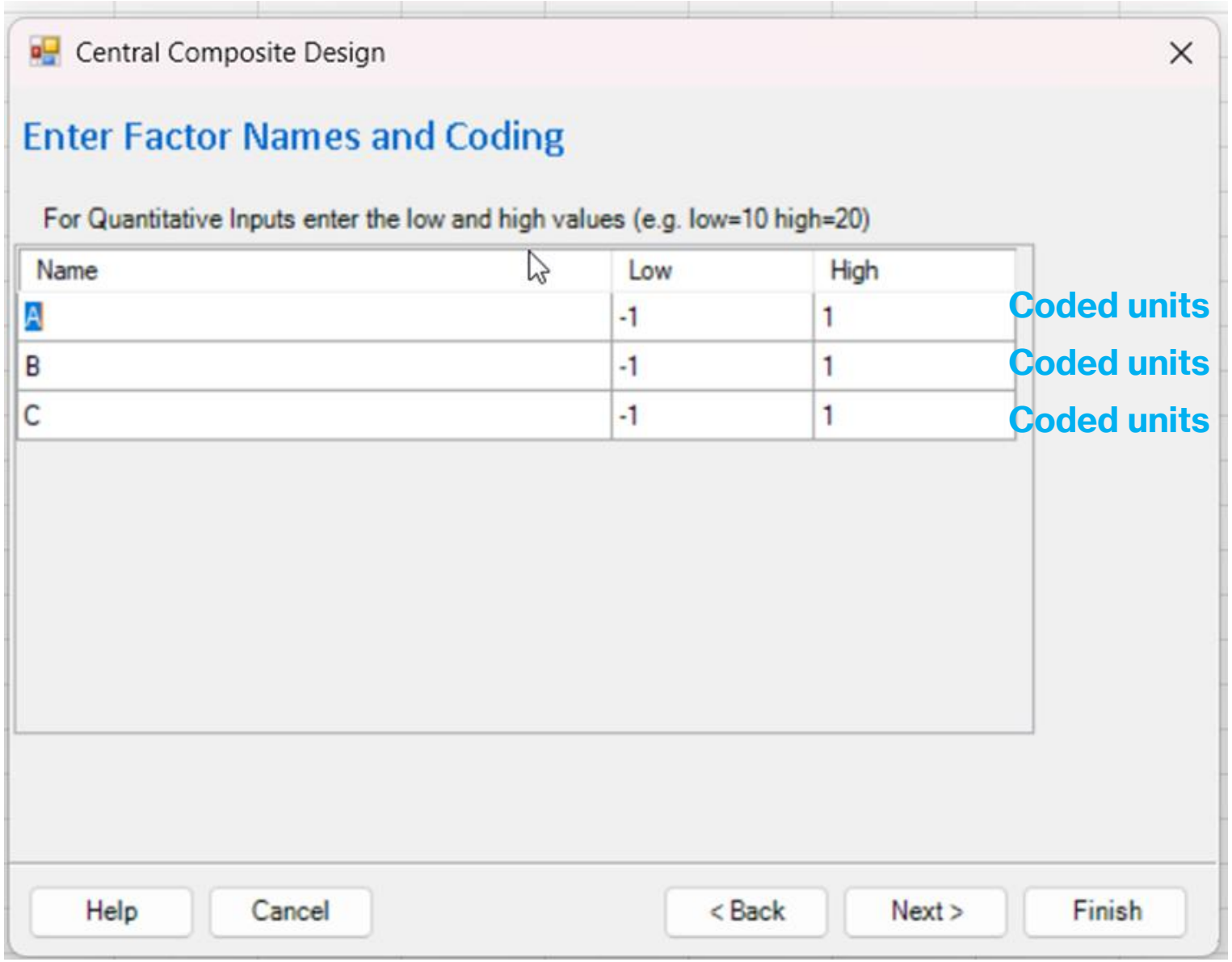
Design Type: Face-Centered Central Composite Design (CCD)

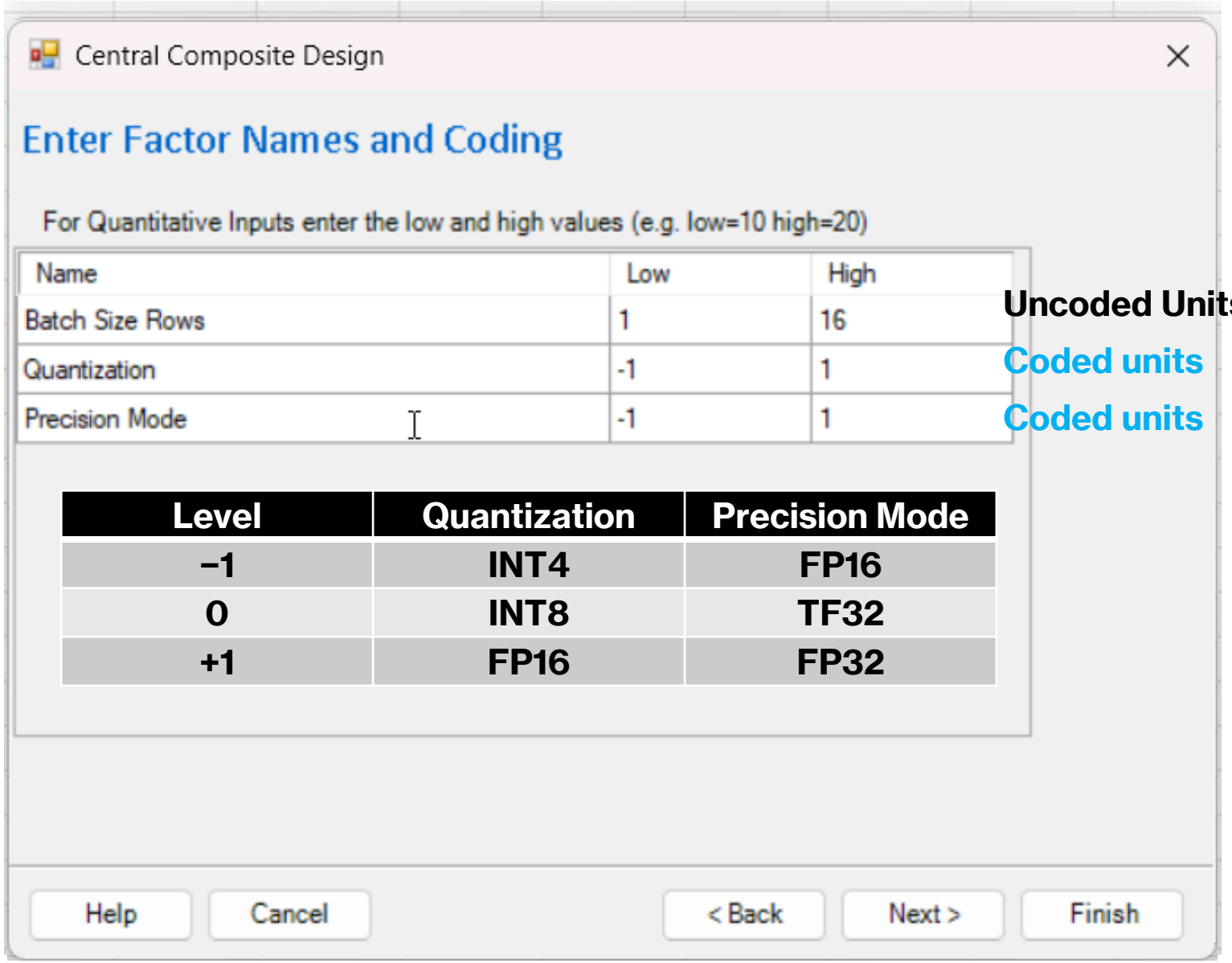
- Three factors
- Curvature expected (batch size saturation)
- Strong interactions (precision mode x batch)
- No unsafe extrapolation beyond feasible runtime settings
- Total runs = 20
  - Runs 1-8 are full factorials
  - Runs 9-14 are face-centered axial points
  - Runs 15-20 are center-point replications to estimate noise and drift.


AutoSave Off
Book1 - Excel
Search
File Home Insert Draw Page Layout Formulas Data Review View Automate Help QXL Monte Carlo QXL Stat Tools QXL DOE
Design Wizard
Create Design
Modify Design
Analyze Design
Optimize
Charts
ProTest (Covering Array)
Options
About
Options
Modeling
Create 2-Level Factorial Design
Create 3-Level Factorial Design
Create N-Level Factorial Design
Create CCD Design
Create Box-Behnken Design
Screening
Create Plackett-Burman Design
Create Taguchi Design
Special
Setup Historical Analysis
Create Custom Design
Create D-Optimal Design
Import
Import From DOE PRO XL
Create CCD Design
Create blocked or unblocked Central Composite Design.
Q7
QXL DOE >Create Design >Create CCD Design

Choose Full factorial, 3 Factors in 8 Runs. Resolution: Full




Central Composite Design
Enter Factor Names and Coding
For Quantitative Inputs enter the low and high values (e.g. low=10 high=20)
Name
Low
High
A
-1
1
B
-1
1
C
-1
1
Coded units
Coded units
Coded units
Help
Cancel
< Back
Next >
Finish

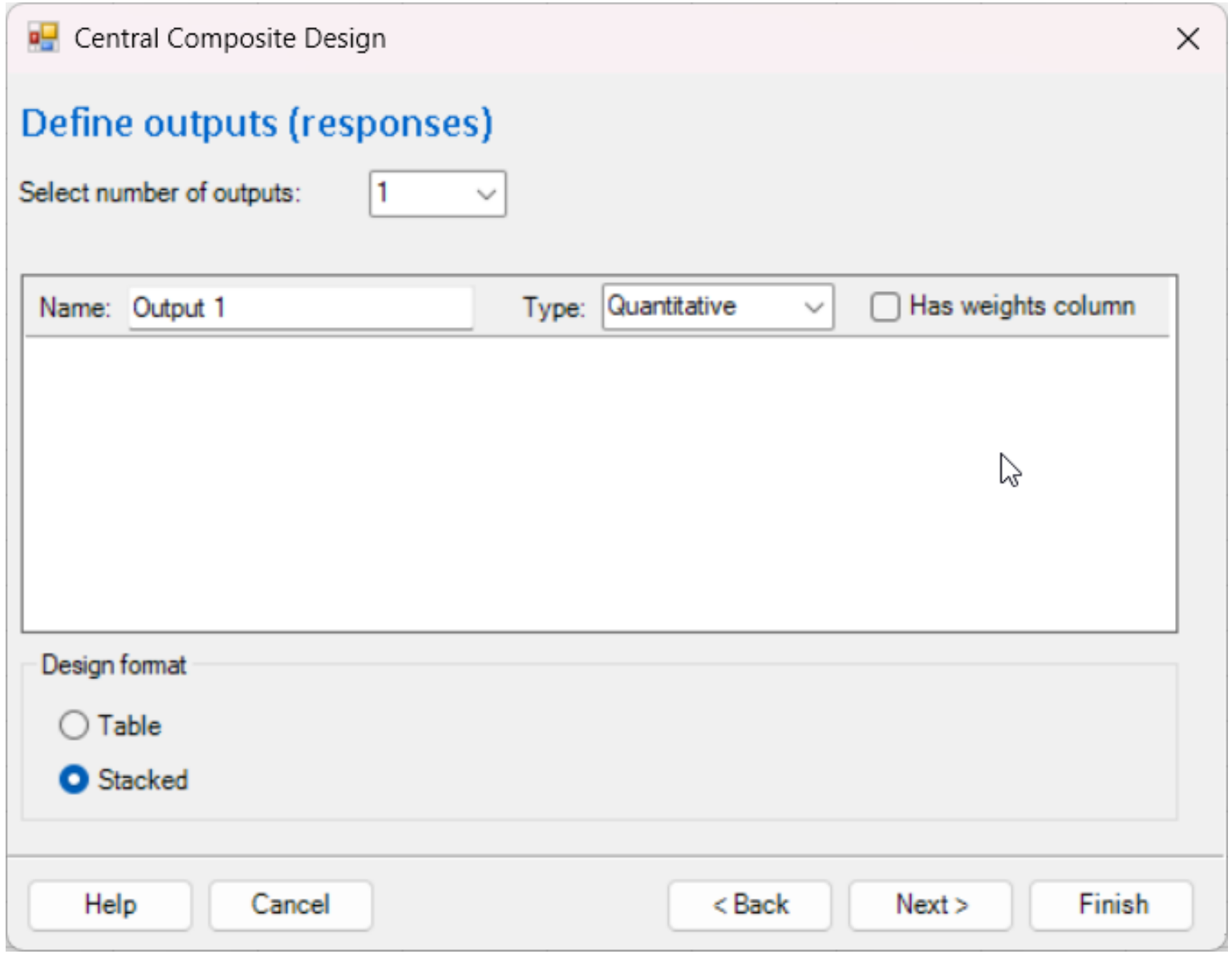


Uncoded Units

Coded units

Coded units

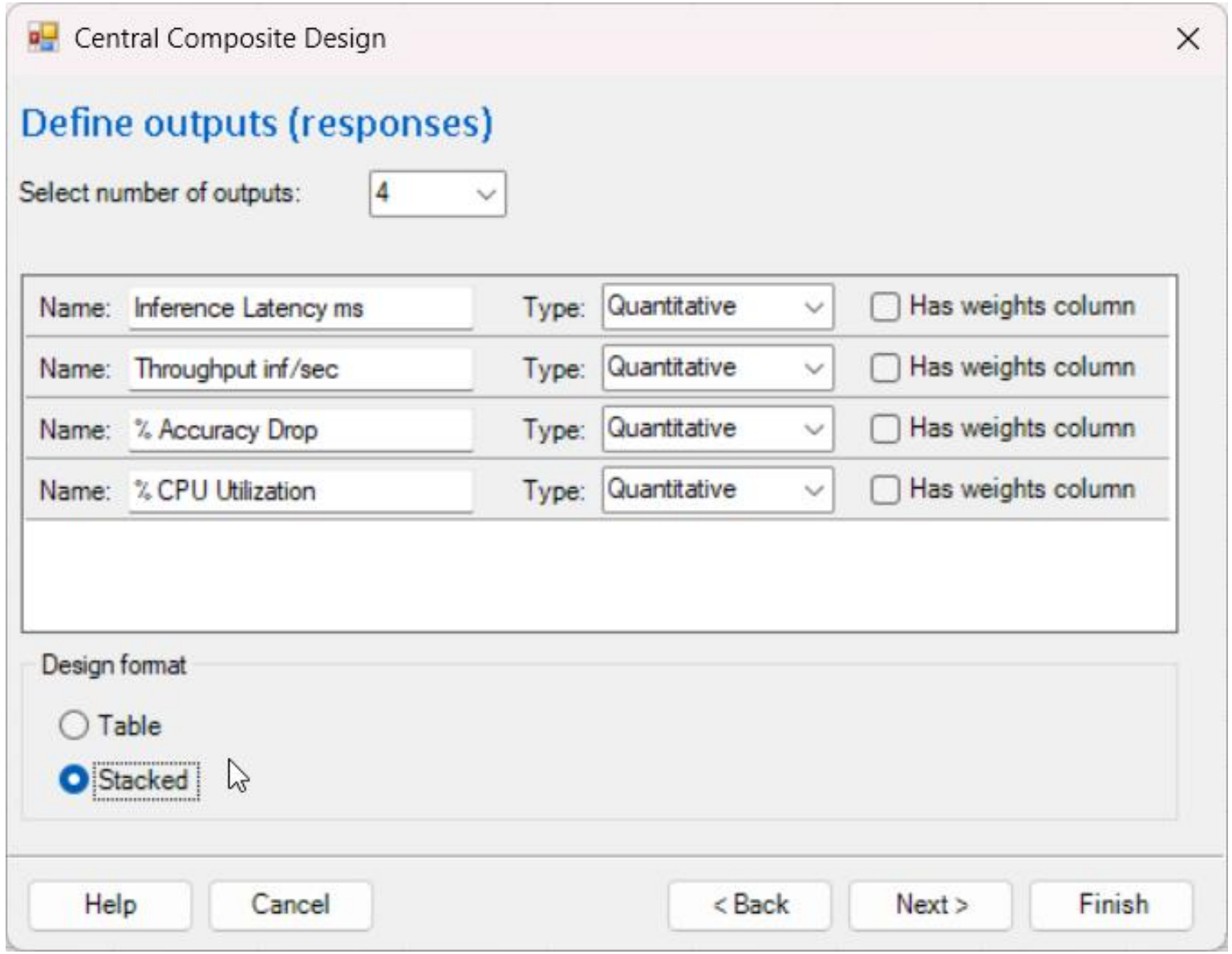
Level	Quantization	Precision Mode
–1	INT4	FP16
0	INT8	TF32
+1	FP16	FP32

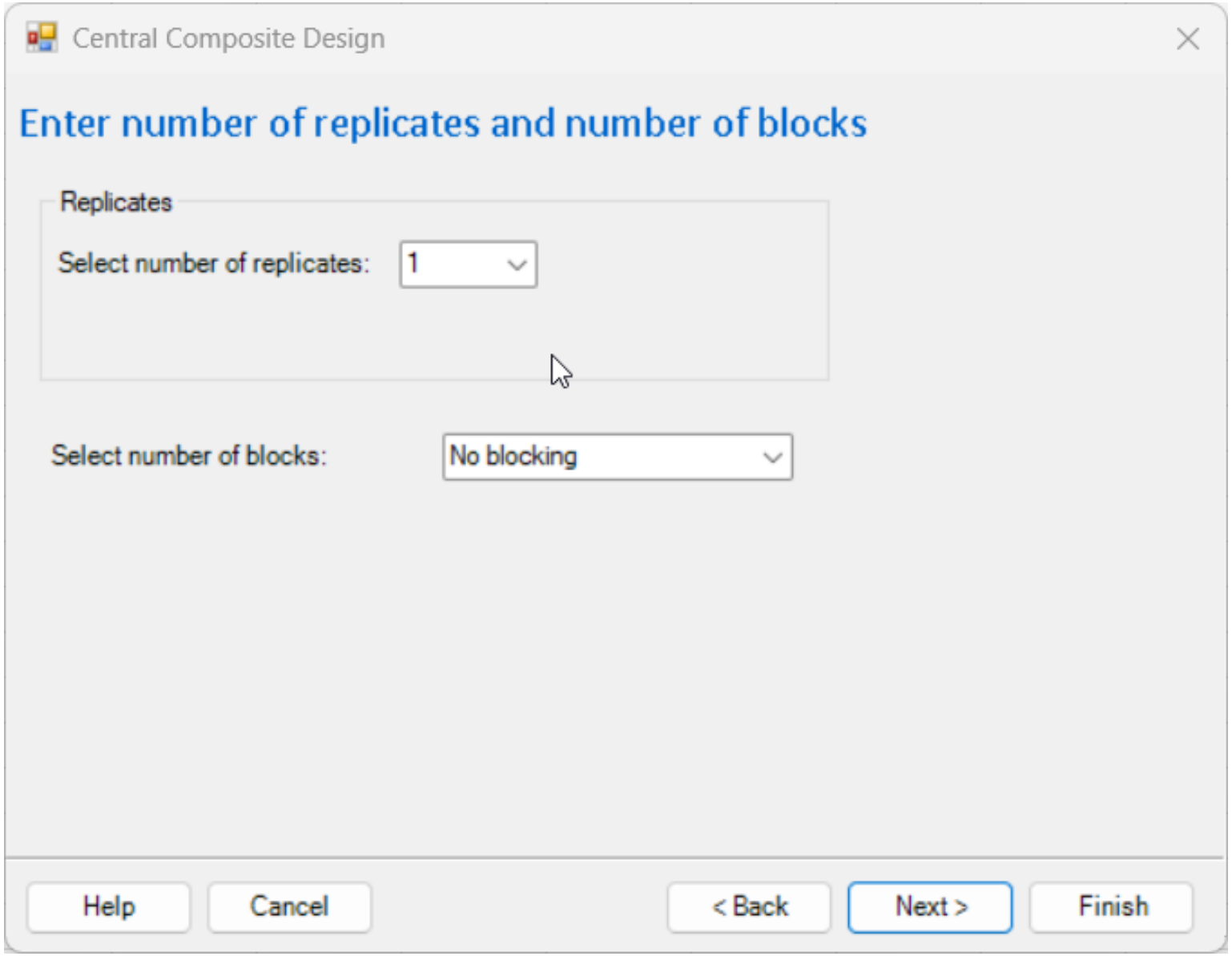

Central Composite Design
Define outputs (responses)
Select number of outputs:
1
Name:
Output 1
Type:
Quantitative
Has weights column
Design format
Table
Stacked
Help
Cancel
< Back
Next >
Finish

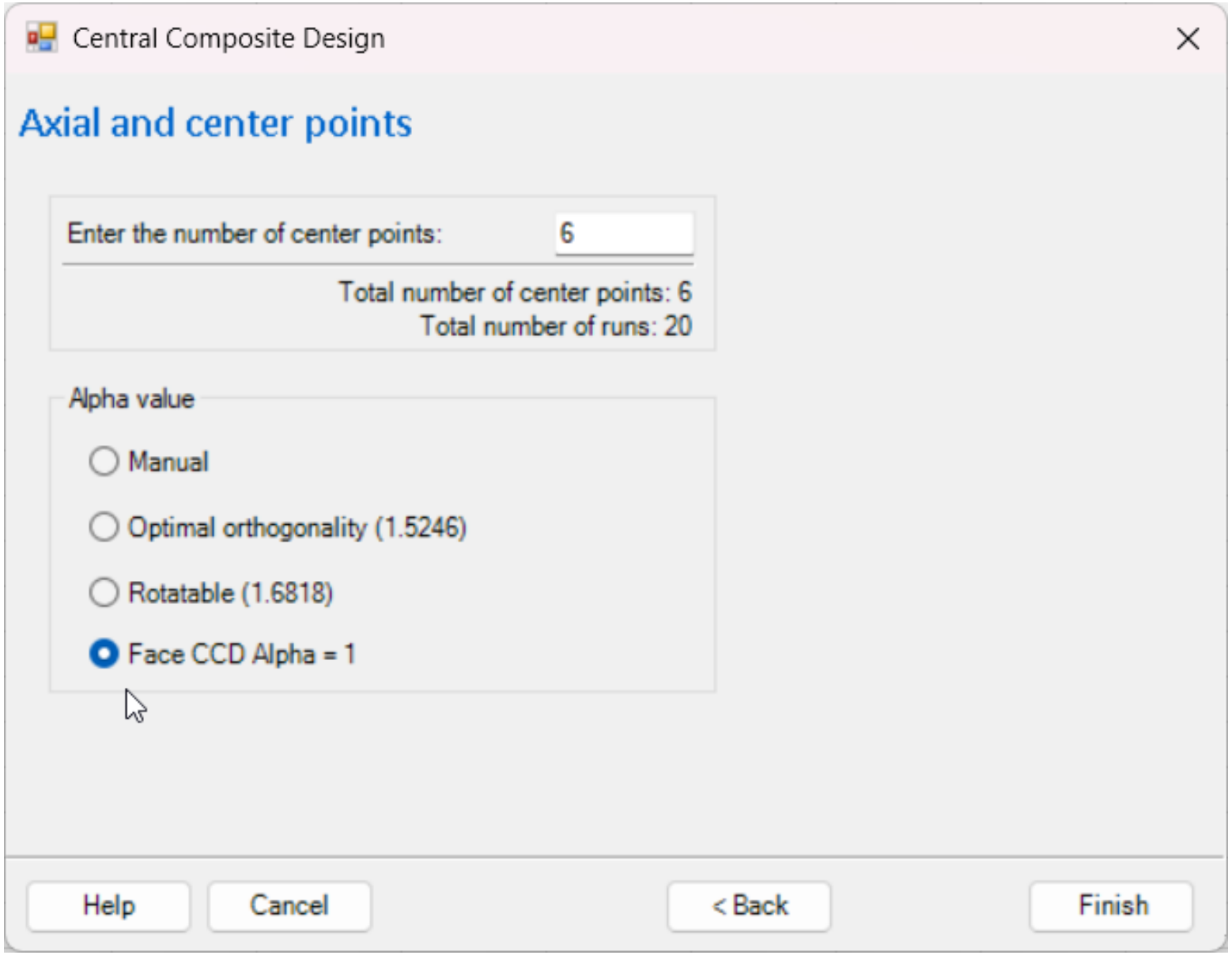
Response Variables

Measure using repeated inference runs under identical input conditions.

Response	Measurement Goal	Specification Limits
Y1 = Inference Latency	**Minimize (milliseconds)**	**USL = 11 ms**
Y2 = Throughput	**Maximize (inferences per sec)**	**LSL = 125 ips**
Y3 = Accuracy Drop	**Minimize (percent)**	**USL = 2%**
Y4 = Power or CPU Utilization	**Constraint / Minimize**	**USL = 68%**


Central Composite Design
Define outputs (responses)
Select number of outputs:
4
Name: Inference Latency ms
Type: Quantitative
Has weights column
Name: Throughput inf/sec
Type: Quantitative
Has weights column
Name: % Accuracy Drop
Type: Quantitative
Has weights column
Name: % CPU Utilization
Type: Quantitative
Has weights column
Design format
Table
Stacked
Help
Cancel
< Back
Next >
Finish


Central Composite Design
Enter number of replicates and number of blocks
Replicates
Select number of replicates:
1
Select number of blocks:
No blocking
Help
Cancel
< Back
Next >
Finish


Central Composite Design
Axial and center points
Enter the number of center points:
6
Total number of center points: 6
Total number of runs: 20
Alpha value
Manual
Optimal orthogonality (1.5246)
Rotatable (1.6818)
Face CCD Alpha = 1
Help
Cancel
< Back
Finish

Batch Size should be an integer; therefore, round down 8.5 to 8.

	A	B	C
Run	Batch Size Rows	Quantization	Precision Mode
1	1	-1	-1
2	1	-1	1
3	1	1	-1
4	1	1	1
5	16	-1	-1
6	16	-1	1
7	16	1	-1
8	16	1	1
9	1	0	0
10	16	0	0
11	8.5	-1	0
12	8.5	1	0
13	8.5	0	-1
14	8.5	0	1
15	8.5	0	0
16	8.5	0	0
17	8.5	0	0
18	8.5	0	0
19	8.5	0	0
20	8.5	0	0

Before

	A	B	C
Run	Batch Size Rows	Quantization	Precision Mode
1	1	-1	-1
2	1	-1	1
3	1	1	-1
4	1	1	1
5	16	-1	-1
6	16	-1	1
7	16	1	-1
8	16	1	1
9	1	0	0
10	16	0	0
11	8	-1	0
12	8	1	0
13	8	0	-1
14	8	0	1
15	8	0	0
16	8	0	0
17	8	0	0
18	8	0	0
19	8	0	0
20	8	0	0

After

Enter Specification Limits.

Inference Latency ms	
USL	11
LSL	
Data	

Throughput inf/sec	
USL	
LSL	125
Data	

% Accuracy Drop	
USL	2
LSL	
Data	

% CPU Utilization	
USL	68
LSL	
Data	

Run experiments and enter response variables.

Inference Latency ms	
USL	11
LSL	
Data	
12	
18	
25	
30	
7	
11	
15	
20	
14	
8	
10	
17	
20	
16	
12	
13	
11	
12	
12	
13	

Throughput inf/sec	
USL	
LSL	125
Data	
80	
60	
40	
35	
180	
140	
95	
85	
70	
165	
90	
65	
60	
75	
110	
108	
112	
109	
111	
107	

% Accuracy Drop	
USL	2
LSL	
Data	
2.8	
0.3	
0.6	
0.1	
3.2	
0.5	
0.9	
0.2	
1.5	
1.9	
2.1	
0.7	
1.1	
0.4	
1.2	
1.1	
1.3	
1.2	
1.1	
1.2	

% CPU Utilization	
USL	68
LSL	
Data	
35	
55	
42	
70	
68	
88	
75	
95	
48	
82	
50	
60	
55	
72	
60	
62	
59	
61	
60	
63	

Quantum XL

Central Composite Design

Base design: Full factorial

3 Factors in 20 Runs

6 Center points. Alpha = 1

Design is not replicated.

	A	B	C
Run	Batch Size Rows	Quantization	Precision Mode
1	1	-1	-1
2	1	-1	1
3	1	1	-1
4	1	1	1
5	16	-1	-1
6	16	-1	1
7	16	1	-1
8	16	1	1
9	1	0	0
10	16	0	0
11	8	-1	0
12	8	1	0
13	8	0	-1
14	8	0	1
15	8	0	0
16	8	0	0
17	8	0	0
18	8	0	0
19	8	0	0
20	8	0	0

edit response name (optional) -->

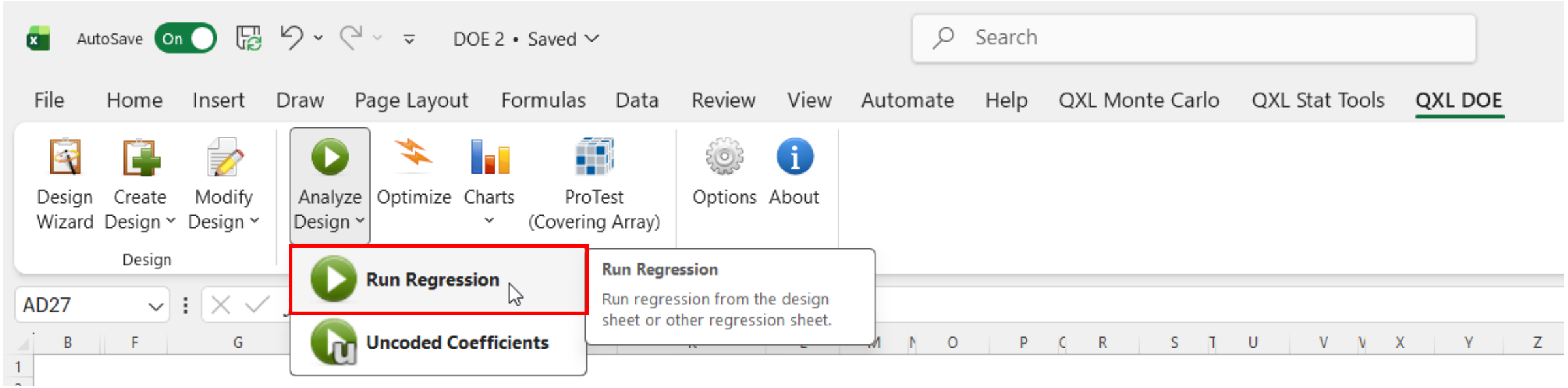
enter spec limits (optional) -->

<-- enter factor names

enter response data here -->

Inference Latency ms		Throughput inf/sec		% Accuracy Drop		% CPU Utilization	
USL	11	USL		USL	2	USL	68
LSL		LSL	125	LSL		LSL	
Data		Data		Data		Data	
12		80		2.8		35	
18		60		0.3		55	
25		40		0.6		42	
30		35		0.1		70	
7		180		3.2		68	
11		140		0.5		88	
15		95		0.9		75	
20		85		0.2		95	
14		70		1.5		48	
8		165		1.9		82	
10		90		2.1		50	
17		65		0.7		60	
20		60		1.1		55	
16		75		0.4		72	
12		110		1.2		60	
13		108		1.1		62	
11		112		1.3		59	
12		109		1.2		61	
12		111		1.1		60	
13		107		1.2		63	

QXL DOE >Analyze Design >Run Regression



1 of 4 Regression Models in Coded Units

	Inference Latency ms						
	Y-Hat						
Name	Factor	Coeff	SE	T	P	VIF	In Model
	Const	12.052	0.7514	16.039	0.000		
Batch Size Rows	Batch Size Rows (A)	-3.8	0.6887	-5.5177	0.000	1.0022	☑
Quantization	Quantization (B)	4.887	0.6888	7.0952	0.000	1.0002	☑
Precision Mode	Precision Mode (C)	1.5975	0.6888	2.3193	0.046	1.0002	☑
Batch Size Rows*Quantization	AB	-0.9758	0.7696	-1.2679	0.237	1.0002	☑
Batch Size Rows*Precision Mode	AC	-0.1898	0.7696	-0.2466	0.811	1.0002	☑
Quantization*Precision Mode	BC	0.0	0.77	0.0	1.000	1.0	☑
Batch Size Rows*Quantization*Precision Mode	ABC	0.25	0.77	0.3247	0.753	1.0	☑
Batch Size Rows*Batch Size Rows	AA	-1.2522	1.32	-0.9487	0.368	1.8204	☑
Quantization*Quantization	BB	1.0	1.3133	0.7614	0.466	1.8182	☑
Precision Mode*Precision Mode	CC	5.5	1.3133	4.1879	0.002	1.8182	☑

Hypothesis Test
Ho: Coefficient is zero
Ha: Coefficient is not zero (term is significant)
Accept Ho if P-value > 0.10 (uncheck "In Model" box)

2 of 4 Regression Models in Coded Units

Name	Throughput inf/sec Y-Hat Factor	Coeff	SE	T	P	VIF	In Model
	Const	105.262	4.9538	21.249	0.000		
Batch Size Rows	Batch Size Rows (A)	38	4.5403	8.3695	0.000	1.0022	☑
Quantization	Quantization (B)	-23.127	4.5408	-5.0932	0.001	1.0002	☑
Precision Mode	Precision Mode (C)	-6.0446	4.5408	-1.3312	0.216	1.0002	☑
Batch Size Rows*Quantization	AB	-9.5415	5.074	-1.8805	0.093	1.0002	☑
Batch Size Rows*Precision Mode	AC	-3.347	5.074	-0.6596	0.526	1.0002	☑
Quantization*Precision Mode	BC	5.625	5.0762	1.1081	0.297	1.0	☑
Batch Size Rows*Quantization*Precision Mode	ABC	1.875	5.0762	0.3694	0.720	1.0	☑
Batch Size Rows*Batch Size Rows	AA	22.247	8.702	2.5566	0.031	1.8204	☑
Quantization*Quantization	BB	-15.318	8.658	-1.7693	0.111	1.8182	☑
Precision Mode*Precision Mode	CC	-25.318	8.658	-2.9243	0.017	1.8182	☑

Hypothesis Test
Ho: Coefficient is zero
Ha: Coefficient is not zero (term is significant)
Accept Ho if P-value > 0.10 (uncheck "In Model" box)

3 of 4 Regression Models in Coded Units

Name
Batch Size Rows
Quantization
Precision Mode
Batch Size Rows*Quantization
Batch Size Rows*Precision Mode
Quantization*Precision Mode
Batch Size Rows*Quantization*Precision Mode
Batch Size Rows*Batch Size Rows
Quantization*Quantization
Precision Mode*Precision Mode

% Accuracy Drop

Y-Hat

Factor	Coeff	SE	T	P	VIF	In Model
Const	1.2505	0.0762	16.421	**0.000**		
Batch Size Rows (A)	0.14	0.0698	2.0059	**0.076**	1.0022	☑
Quantization (B)	-0.6403	0.0698	-9.1733	**0.000**	1.0002	☑
Precision Mode (C)	-0.7107	0.0698	-10.182	**0.000**	1.0002	☑
AB	-0.024	0.078	-0.3074	0.766	1.0002	☑
AC	-0.056	0.078	-0.7173	0.491	1.0002	☑
BC	0.5	0.078	6.4075	**0.000**	1.0	☑
ABC	0.0	0.078	0.0	1.000	1.0	☑
AA	0.3605	0.1338	2.6946	**0.025**	1.8204	☑
BB	0.0682	0.1331	0.5123	0.621	1.8182	☑
CC	-0.5818	0.1331	-4.3715	**0.002**	1.8182	☑

Hypothesis Test
Ho: Coefficient is zero
Ha: Coefficient is not zero (term is significant)
Accept Ho if P-value > 0.10 (uncheck "In Model" box)

3 of 4 Regression Models in Coded Units

Name
Batch Size Rows
Quantization
Precision Mode
Batch Size Rows*Quantization
Batch Size Rows*Precision Mode
Quantization*Precision Mode
Batch Size Rows*Quantization*Precision Mode
Batch Size Rows*Batch Size Rows
Quantization*Quantization
Precision Mode*Precision Mode

% CPU Utilization

Y-Hat

Factor	Coeff	SE	T	P	VIF	In Model
Const	61.26	0.6591	92.948	**0.000**		
Batch Size Rows (A)	15.8	0.6041	26.156	**0.000**	1.0022	☑
Quantization (B)	4.5866	0.6041	7.592	**0.000**	1.0002	☑
Precision Mode (C)	10.487	0.6041	17.359	**0.000**	1.0002	☑
AB	-1.0058	0.6751	-1.4899	0.170	1.0002	☑
AC	-0.9658	0.6751	-1.4307	0.186	1.0002	☑
BC	1.0	0.6754	1.4807	0.173	1.0	☑
ABC	-1.0	0.6754	-1.4807	0.173	1.0	☑
AA	4.6491	1.1578	4.0157	**0.003**	1.8204	☑
BB	-4.3182	1.1519	-3.7487	**0.005**	1.8182	☑
CC	4.1818	1.1519	3.6304	**0.005**	1.8182	☑

Hypothesis Test
Ho: Coefficient is zero
Ha: Coefficient is not zero (term is significant)
Accept Ho if P-value > 0.10 (uncheck "In Model" box)

1) Uncheck the box for insignificant terms.

2) Reduce the model by running regression.

	Inference Latency ms						
	Y-Hat						
Name	Factor	Coeff	SE	T	P	VIF	In Model
	Const	11.996	0.6134	19.557	0.000		
Batch Size Rows	Batch Size Rows (A)	-3.8186	0.6121	-6.2386	0.000	1.0008	☑
Quantization	Quantization (B)	4.9	0.6125	7.9997	0.000	1.0	☑
Precision Mode	Precision Mode (C)	1.6	0.6125	2.6121	0.020	1.0	☑
Batch Size Rows*Quantization	AB						☐
Quantization*Precision Mode	BC						☐
Batch Size Rows*Batch Size Rows	AA						☐
Quantization*Quantization	BB						☐
Precision Mode*Precision Mode	CC	5.3527	0.8666	6.1768	0.000	1.0008	☑
	R^2	0.9067					

1 of 4 Reduced Models in Coded Coefficients. Note the R-squared .

1) Uncheck the box for insignificant terms.

2) Reduce the model by running regression.

	Throughput inf/sec						
	Y-Hat						
Name	Factor	Coeff	SE	T	P	VIF	In Model
	Const	103.373	5.0073	20.644	0.000		
Batch Size Rows	Batch Size Rows (A)	38	4.6998	8.0854	0.000	1.0022	☑
Quantization	Quantization (B)	-23.127	4.7004	-4.9203	0.000	1.0002	☑
Precision Mode	Precision Mode (C)	-6.0	4.6998	-1.2766	0.224	1.0	☑
Batch Size Rows*Quantization	AB	-9.5415	5.2523	-1.8167	0.092	1.0002	☑
Quantization*Precision Mode	BC						☐
Batch Size Rows*Batch Size Rows	AA	16.477	8.3513	1.973	0.070	1.5647	☑
Quantization*Quantization	BB						☐
Precision Mode*Precision Mode	CC	-31.063	8.3082	-3.7388	0.002	1.5625	☑
	R^2	0.8928					

2 of 4 Reduced Models in Coded Coefficients. Note the R-squared .

1) Uncheck the box for insignificant terms.

2) Reduce the model by running regression.

	% Accuracy Drop						
	Y-Hat						
Name	Factor	Coeff	SE	T	P	VIF	In Model
	Const	1.2589	0.0648	19.427	0.000		
Batch Size Rows	Batch Size Rows (A)	0.14	0.0608	2.3019	0.039	1.0022	☑
Quantization	Quantization (B)	-0.64	0.0608	-10.523	0.000	1.0	☑
Precision Mode	Precision Mode (C)	-0.71	0.0608	-11.674	0.000	1.0	☑
Batch Size Rows*Quantization	AB						☐
Quantization*Precision Mode	BC	0.5	0.068	7.3531	0.000	1.0	☑
Batch Size Rows*Batch Size Rows	AA	0.3861	0.1081	3.5729	0.003	1.5647	☑
Quantization*Quantization	BB						☐
Precision Mode*Precision Mode	CC	-0.5563	0.1075	-5.1737	0.000	1.5625	☑
	R^2	0.9625					

3 of 4 Reduced Models in Coded Coefficients. Note the R-squared .

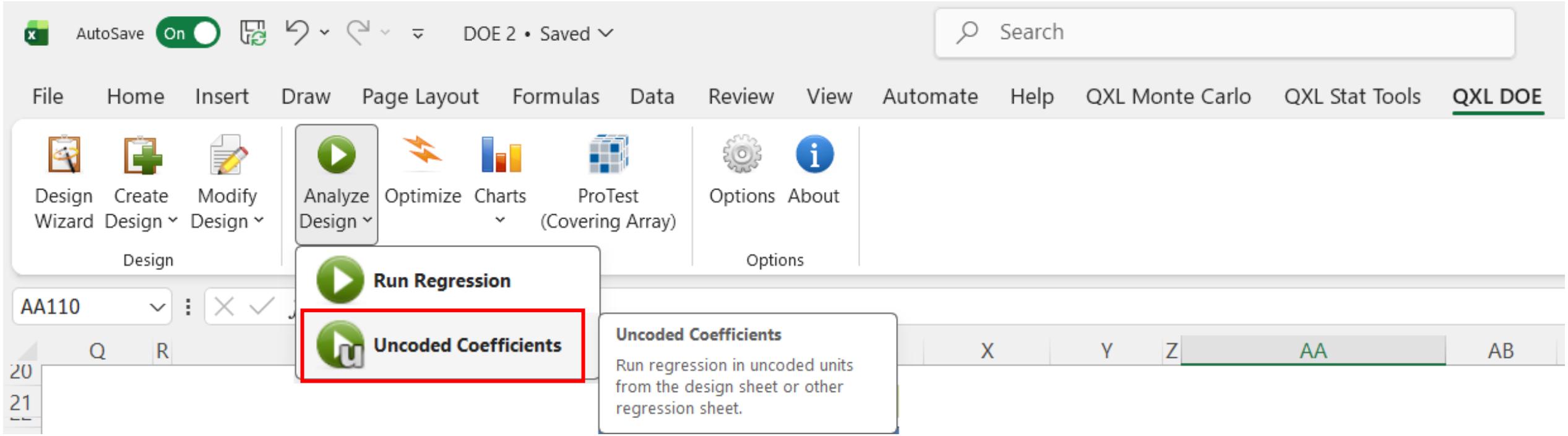
1) Uncheck the box for insignificant terms.

2) Reduce the model by running regression.

Name	% CPU Utilization Y-Hat Factor	Coeff	SE	T	P	VIF	In Model
	Const	61.26	0.768	79.767	0.000		
Batch Size Rows	Batch Size Rows (A)	15.8	0.7039	22.447	0.000	1.0022	☑
Quantization	Quantization (B)	4.6	0.7039	6.5352	0.000	1.0	☑
Precision Mode	Precision Mode (C)	10.5	0.7039	14.917	0.000	1.0	☑
Batch Size Rows*Quantization	AB						☐
Quantization*Precision Mode	BC						☐
Batch Size Rows*Batch Size Rows	AA	4.6491	1.3491	3.4462	0.004	1.8204	☑
Quantization*Quantization	BB	-4.3182	1.3423	-3.2171	0.007	1.8182	☑
Precision Mode*Precision Mode	CC	4.1818	1.3423	3.1155	0.008	1.8182	☑
	R^2	0.9843					

4 of 4 Reduced Models in Coded Coefficients. Note the R-squared .

QXL DOE >Analyze Design >Uncoded Coefficients



1 of 4 Regression Models in Uncoded Coefficients

	Inference Latency ms						
	Y-Hat						
Name	Factor	Coeff	SE	T	P	VIF	In Model
	Const	16.324	0.9012	18.113			
Batch Size Rows	Batch Size Rows (A)	-0.5091	0.0816	-6.2386		1.0008	☑
Quantization	Quantization (B)	4.9	0.6125	7.9997		1.0	☑
Precision Mode	Precision Mode (C)	1.6	0.6125	2.6121		1.0	☑
Batch Size Rows*Quantization	AB						☐
Quantization*Precision Mode	BC						☐
Batch Size Rows*Batch Size Rows	AA						☐
Quantization*Quantization	BB						☐
Precision Mode*Precision Mode	CC	5.3527	0.8666	6.1768		1.0008	☑
	R^2	0.9067					

2 of 4 Regression Models in Uncoded Coefficients

	Throughput inf/sec						
	Y-Hat						
Name	**Factor**	**Coeff**	**SE**	**T**	**P**	**VIF**	**In Model**
	Const	81.47	11.523	7.07			
Batch Size Rows	**Batch Size Rows (A)**	0.0868	2.6234	0.0331		17.565	☑
Quantization	**Quantization (B)**	-12.313	7.5295	-1.6354		2.5666	☑
Precision Mode	**Precision Mode (C)**	-6.0	4.6998	-1.2766		1.0	☑
Batch Size Rows*Quantization	**AB**	-1.2722	0.7003	-1.8167		2.5666	☑
Quantization*Precision Mode	**BC**						☐
Batch Size Rows*Batch Size Rows	**AA**	0.2929	0.1485	1.973		18.298	☑
Quantization*Quantization	**BB**						☐
Precision Mode*Precision Mode	**CC**	-31.063	8.3082	-3.7388		1.5625	☑
	R^2	0.8928					

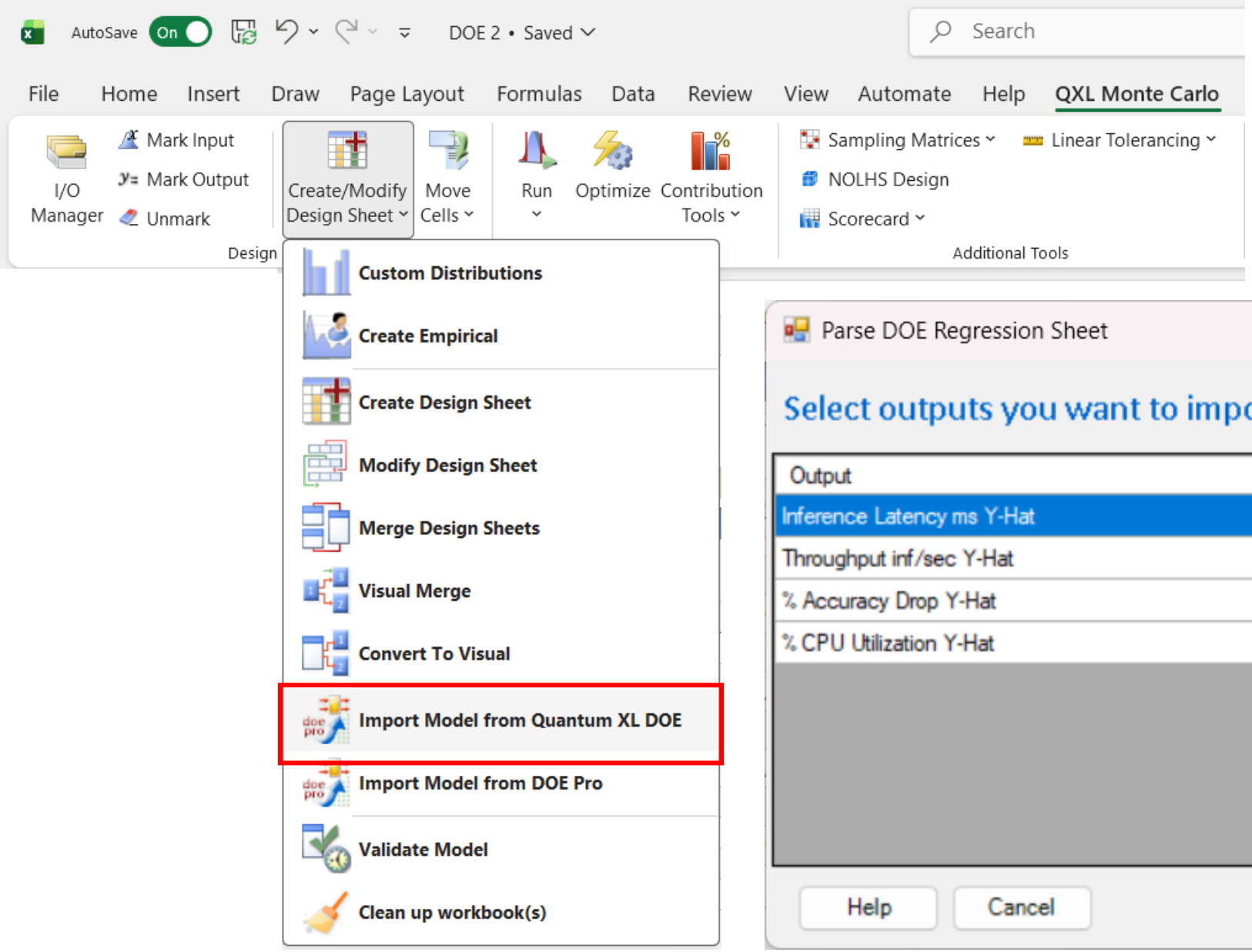
3 of 4 Regression Models in Uncoded Coefficients

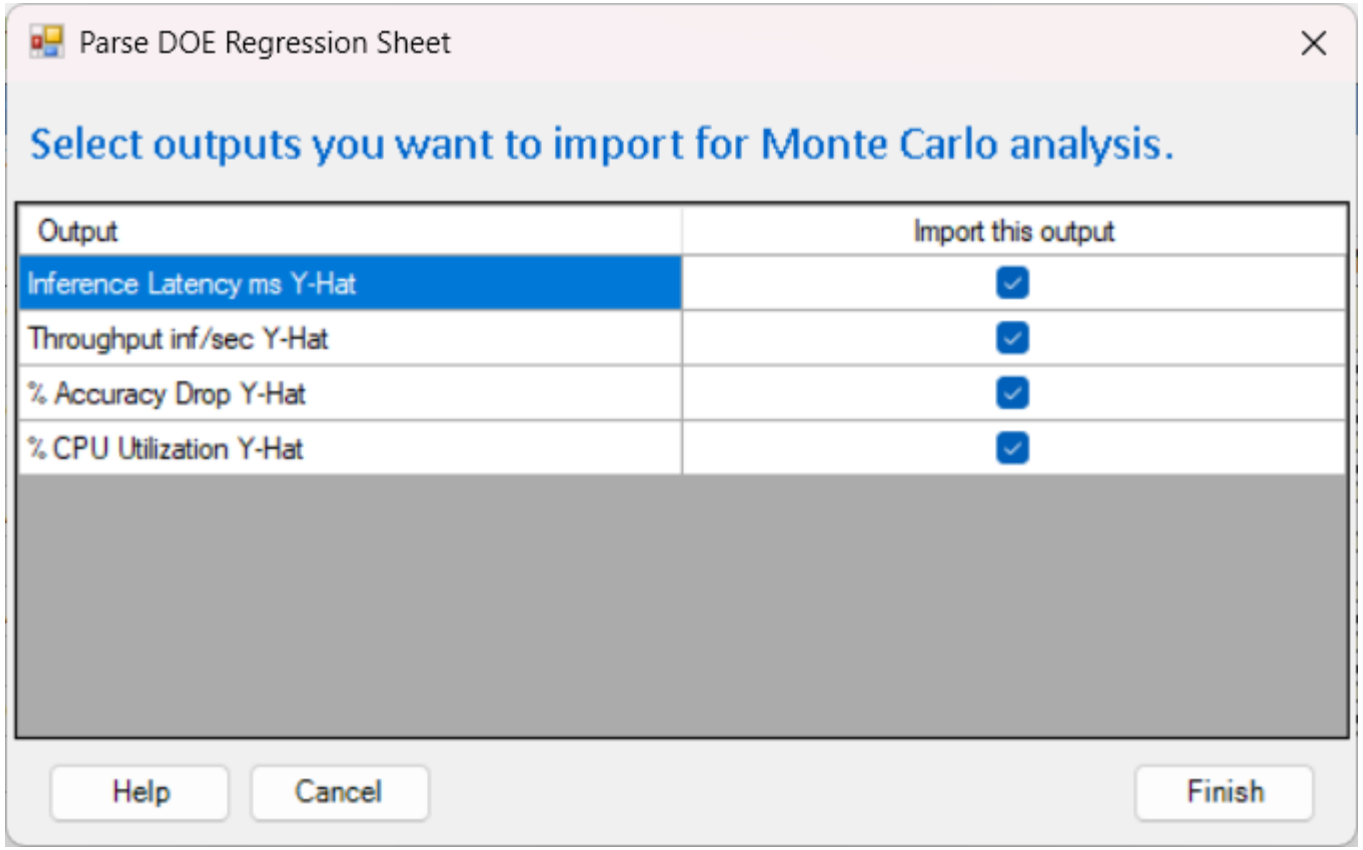
	% Accuracy Drop						
	Y-Hat						
Name	Factor	Coeff	SE	T	P	VIF	In Model
	Const	1.5962	0.1491	10.704			
Batch Size Rows	Batch Size Rows (A)	-0.098	0.0339	-2.8876		17.565	☑
Quantization	Quantization (B)	-0.64	0.0608	-10.523		1.0	☑
Precision Mode	Precision Mode (C)	-0.71	0.0608	-11.674		1.0	☑
Batch Size Rows*Quantization	AB						☐
Quantization*Precision Mode	BC	0.5	0.068	7.3531		1.0	☑
Batch Size Rows*Batch Size Rows	AA	0.0069	0.0019	3.5729		18.298	☑
Quantization*Quantization	BB						☐
Precision Mode*Precision Mode	CC	-0.5563	0.1075	-5.1737		1.5625	☑
	R^2	0.9625					

4 of 4 Regression Models in Uncoded Coefficients

Name	% CPU Utilization Y-Hat Factor	Coeff	SE	T	P	VIF	In Model
	Const	49.325	1.9086	25.844			
Batch Size Rows	**Batch Size Rows (A)**	0.7016	0.4216	1.6643		20.221	☑
Quantization	**Quantization (B)**	4.6	0.7039	6.5352		1.0	☑
Precision Mode	**Precision Mode (C)**	10.5	0.7039	14.917		1.0	☑
Batch Size Rows*Quantization	**AB**						☐
Quantization*Precision Mode	**BC**						☐
Batch Size Rows*Batch Size Rows	**AA**	0.0827	0.024	3.4462		21.288	☑
Quantization*Quantization	**BB**	-4.3182	1.3423	-3.2171		1.8182	☑
Precision Mode*Precision Mode	**CC**	4.1818	1.3423	3.1155		1.8182	☑
	R^2	0.9843					
	..						

QXL Monte Carlo >Create/Modify Design Sheet >Import Model from Quantum XL DOE





Enter 0 for 2nd Parameter in Monte Carlo Simulation Model

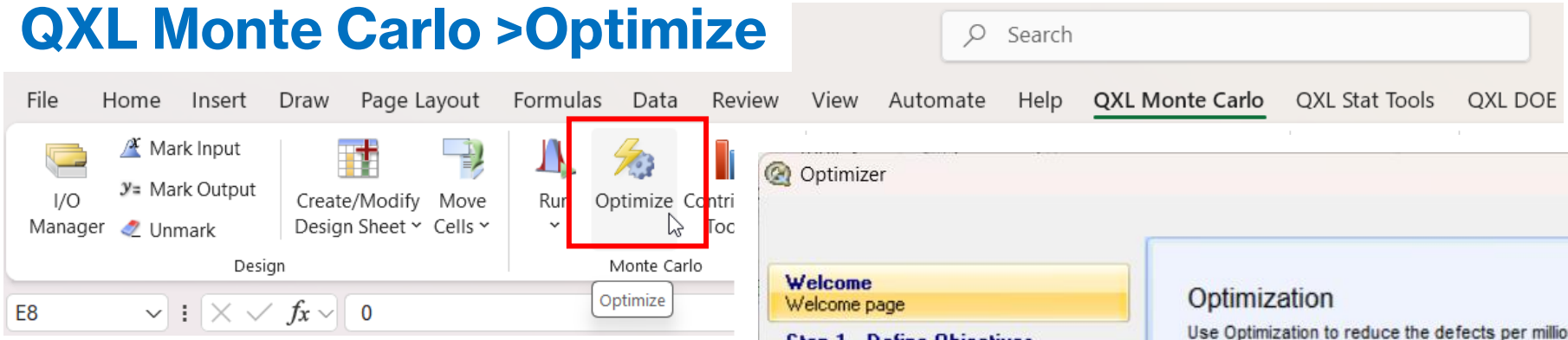
Transfer function(s) in actual units.

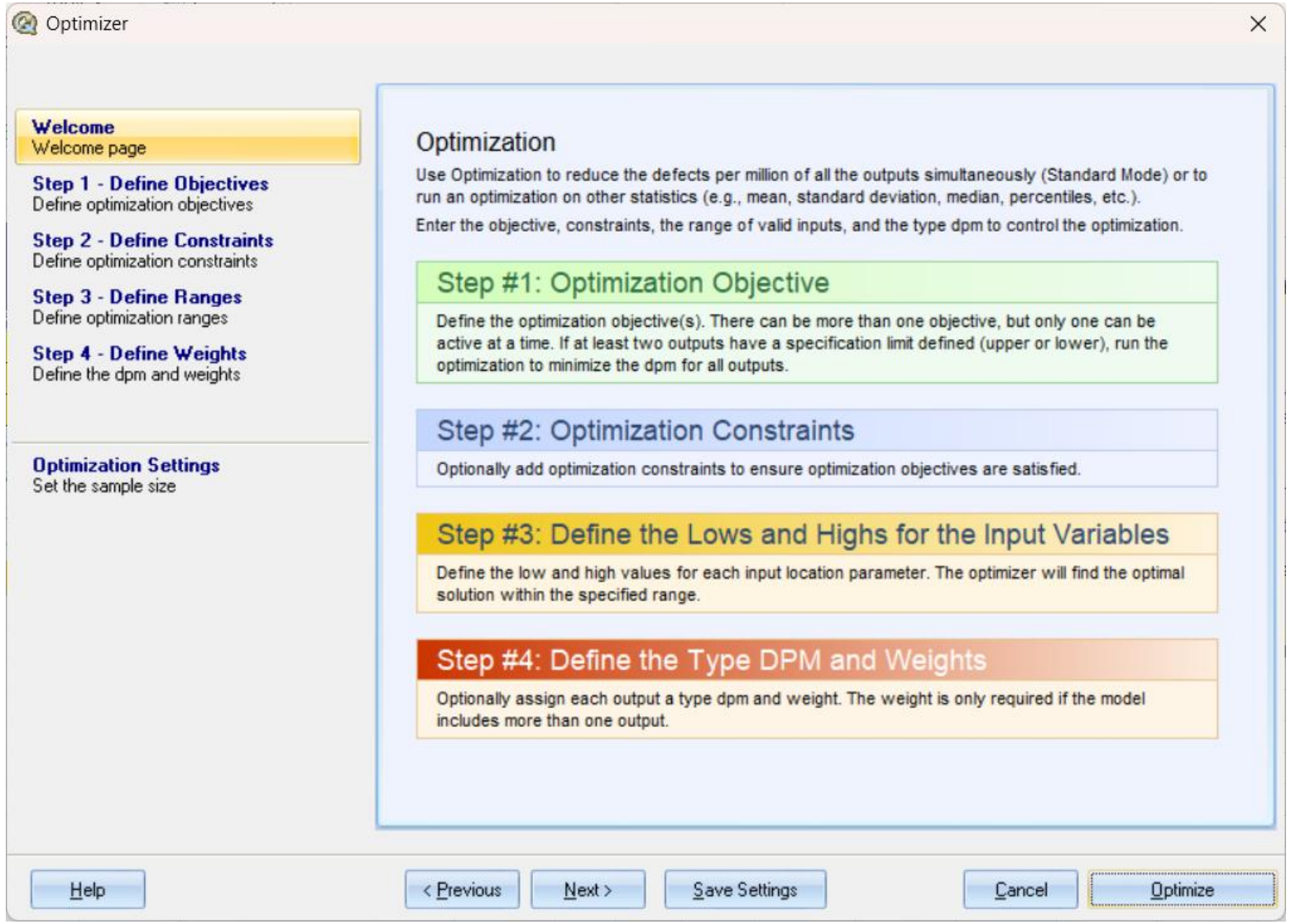
Process Inputs				
Factor	Distro	1st Parameter	2nd Parameter	Exper
Batch Size Rows	Normal	8.5	0	8.5
Quantization	Normal	0	0	0
Precision Mode	Normal	0	0	0
Noise_Inference Latency ms	Normal	0	1.936972136	0
Noise_Throughput inf/sec	Normal	0	14.86222748	0
Noise_% Accuracy Drop	Normal	0	0.192328845	0
Noise_% CPU Utilization	Normal	0	2.225880827	0
Use Quantum XL > Modify Design Sheet to add more inputs				

Process Outputs				
Name	Function	LSL	USL	Active
Inference Latency ms	11.99634		11	X
Throughput inf/sec	103.3726	125		X
% Accuracy Drop	1.258867		2	X
% CPU Utilization	61.25994		68	X
Use Quantum XL > Modify Design Sheet to add more outputs				

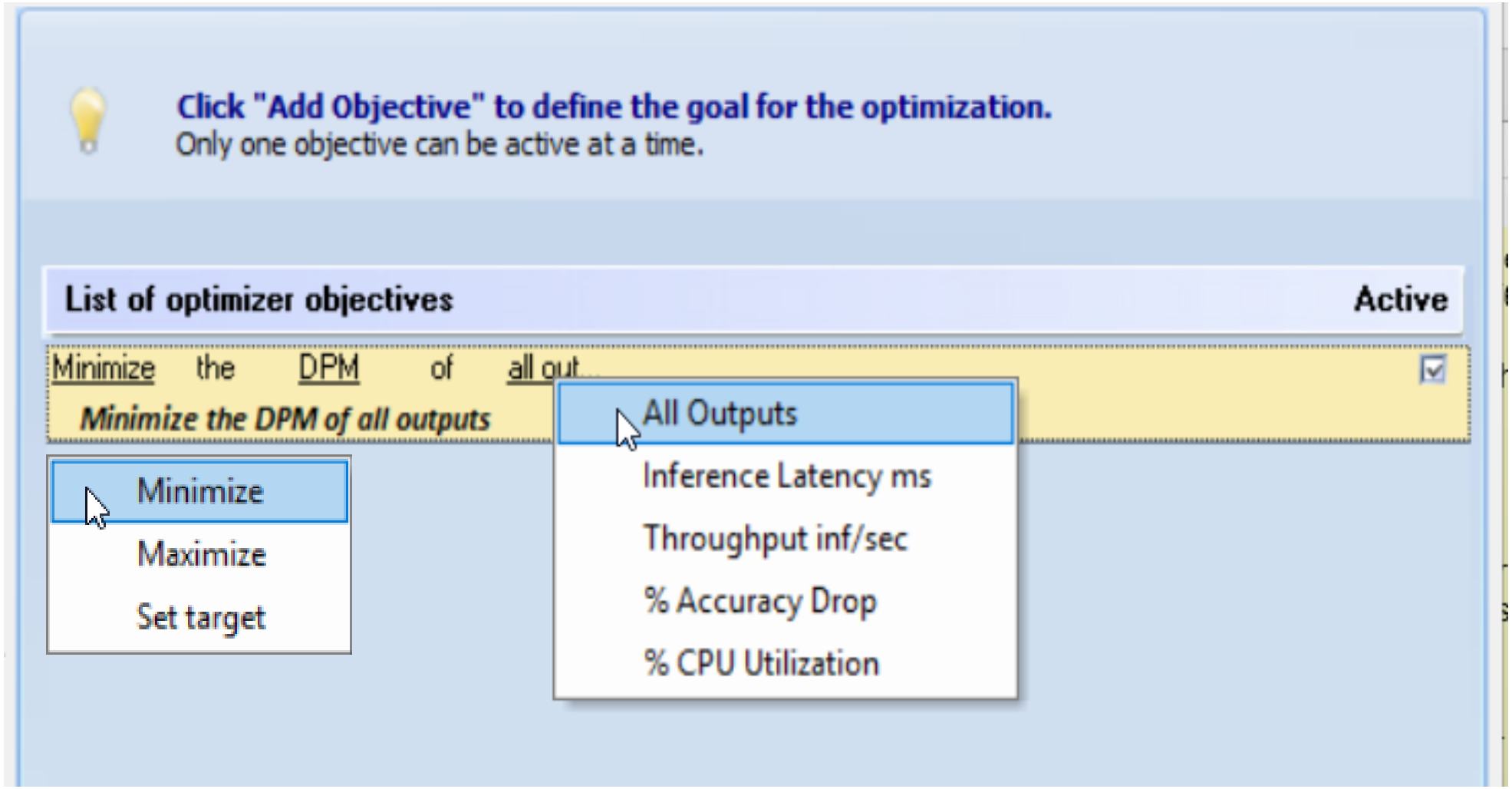
Firmware/Software parameters do not have tolerances.

QXL Monte Carlo >Optimize

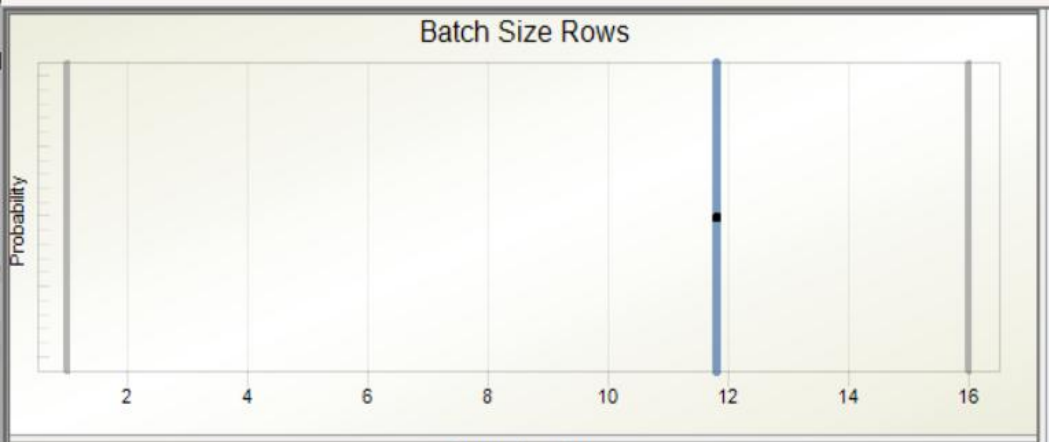


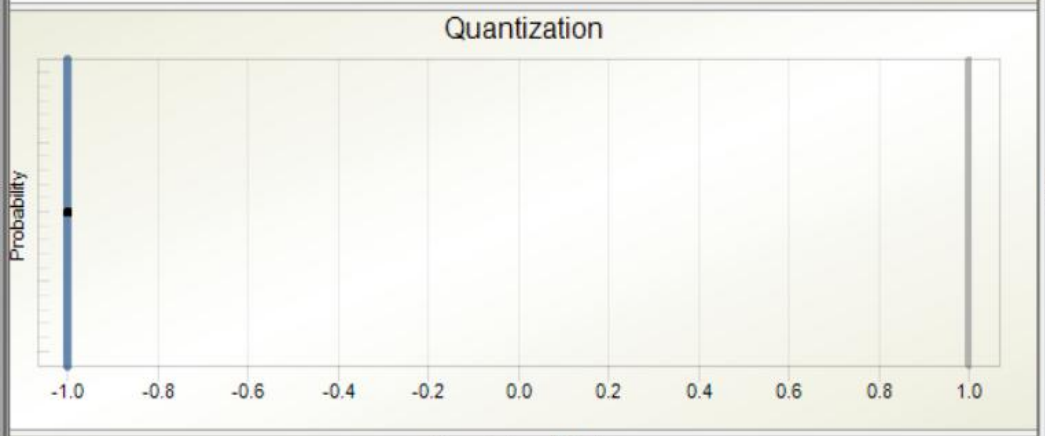


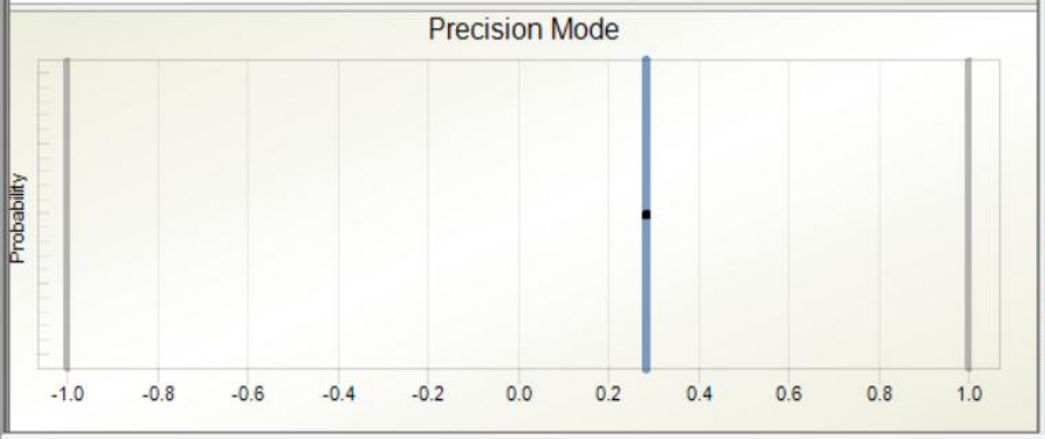
Keep Default Settings for Optimizer Objectives

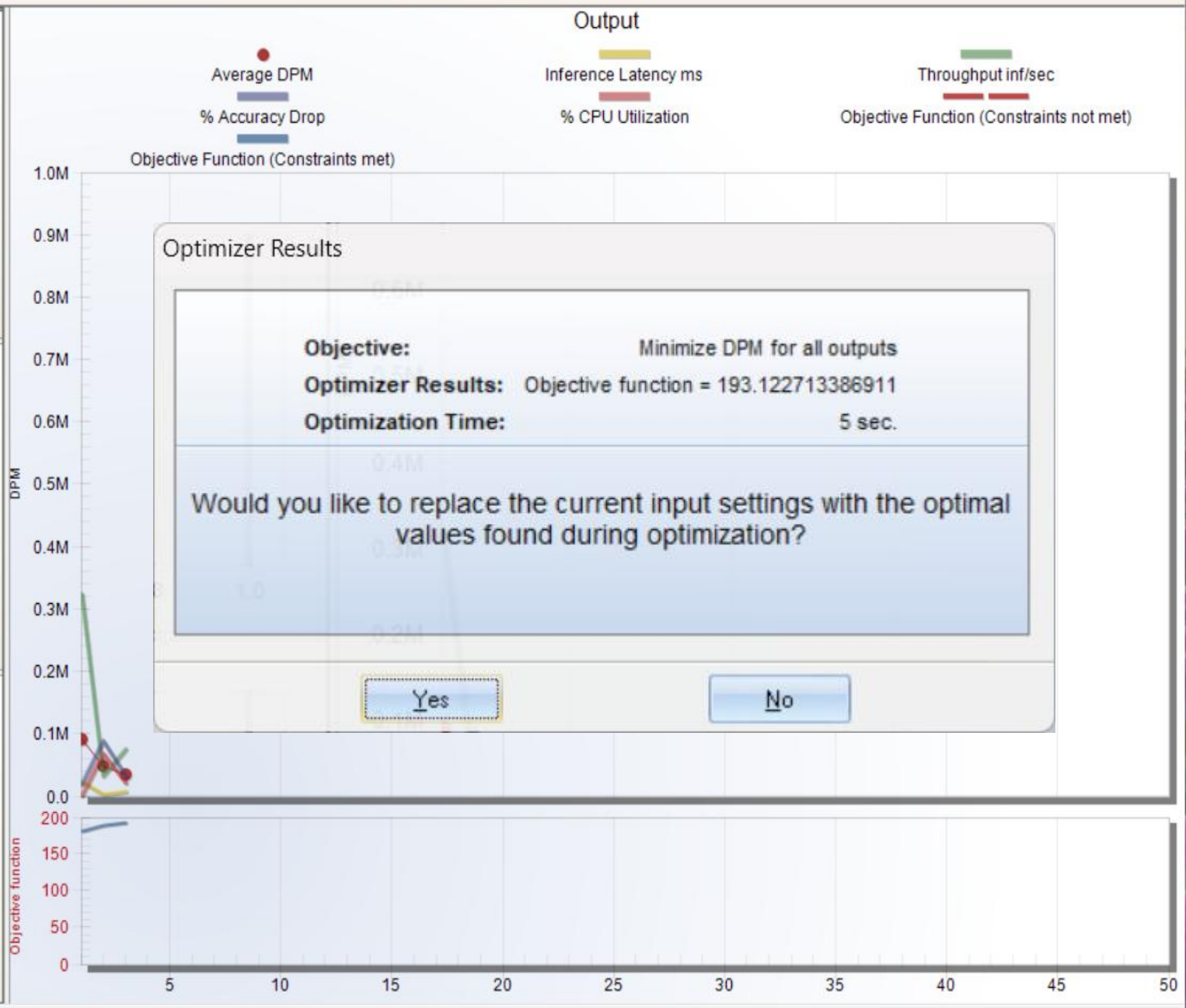


Real-Time Optimization Progress





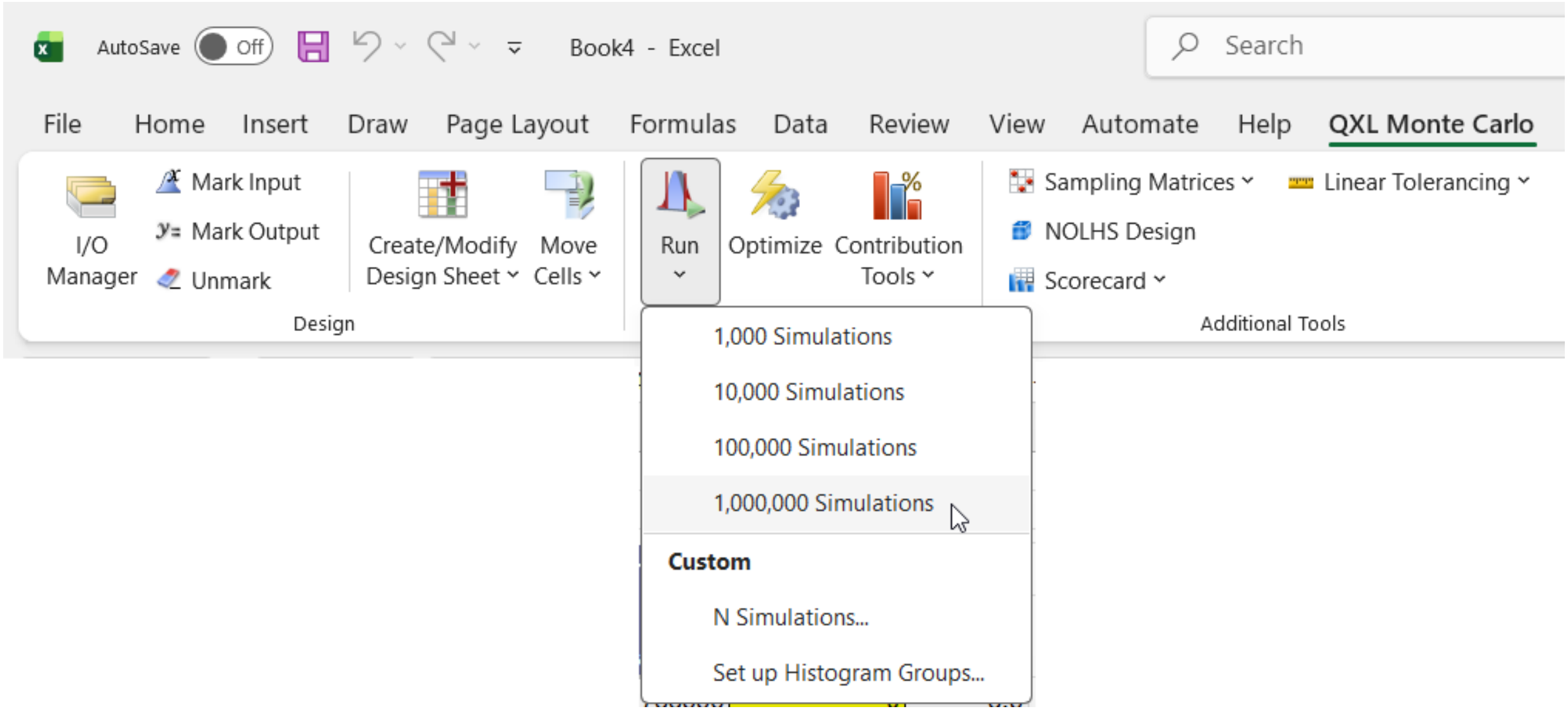




Update graph fast — Optimize fast — Time elapsed: 1 sec.

Optimization will stop automatically when complete. | Stop Optimization and Save Results | Cancel | Pau40

QXL Monte Carlo >Run >1,000,000 Simulations



Converting DPM to Z-score

- Defects Per Million (**DPM**) is a better indicator of quality than Cpk when:
  - The distribution is non-normal or asymmetrical.
  - One is using Monte Carlo simulation to generate a million data points.

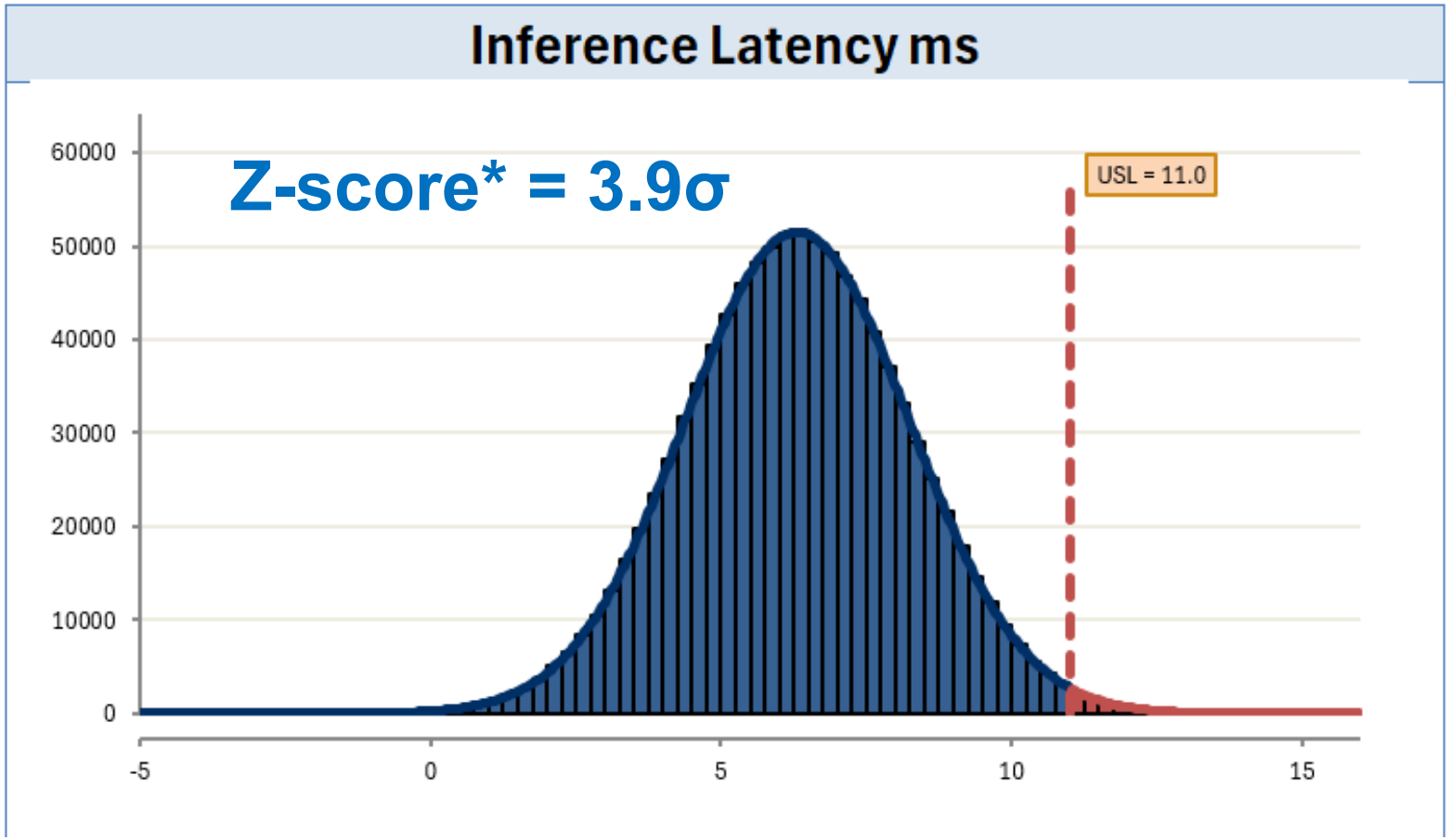
- **Yield = (10^6 – DPM) / 10^6**

- **Sigma Long-term = NORM.S.INV(Yield)**

- **Sigma Short-term or Z-score = Sigma Long-term + 1.5**

Process Inputs		
Batch Size Rows	Mean	StDev
Distribution: **Normal**	12	0
Quantization	Mean	StDev
Distribution: **Normal**	-1	0
Precision Mode	Mean	StDev
Distribution: **Normal**	0	0

Level	Quantization	Precision Mode
−1	INT4	FP16
0	INT8	TF32
+1	FP16	FP32



Process Outputs	
# of Simulations	1,000,000
Mean	6.308709
StDev	1.936854
Median	6.307316
LSL	
USL	11

Normal Distro Statistics	
KS Test p-Value (Normal)	Not Avail
dpm	7,715
Cpk	0.807
Cp	

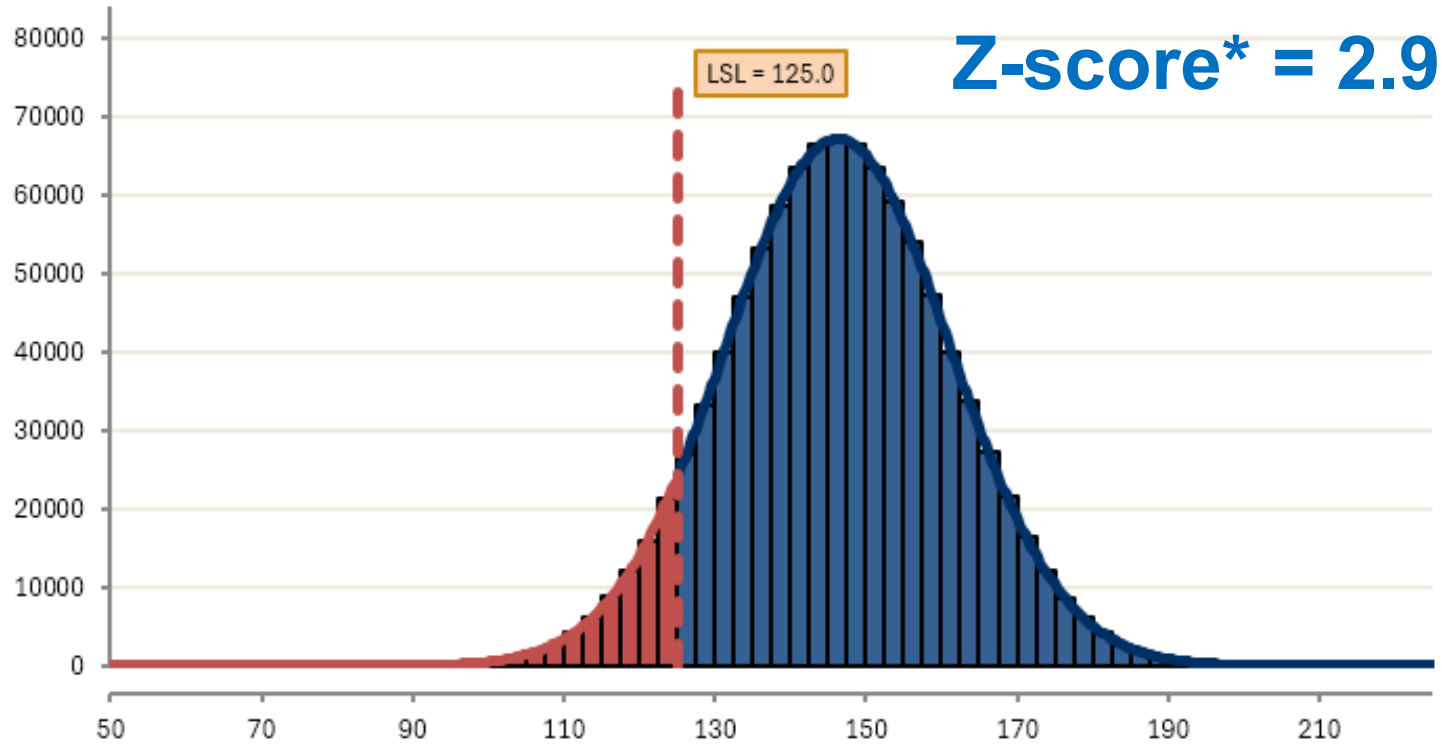
Observed Defect Statistics	
Simulations outside of spec	7,795

*note: based on Observed DPM

Process Inputs		
Batch Size Rows	Mean	StDev
Distribution: **Normal**	12	0
Quantization	Mean	StDev
Distribution: **Normal**	-1	0
Precision Mode	Mean	StDev
Distribution: **Normal**	0	0

Level	Quantization	Precision Mode
−1	INT4	FP16
0	INT8	TF32
+1	FP16	FP32

Throughput inf/sec



Z-score* = 2.9σ

Process Outputs	
# of Simulations	1,000,000
Mean	146.29
StDev	14.86
Median	146.30
LSL	125
USL	

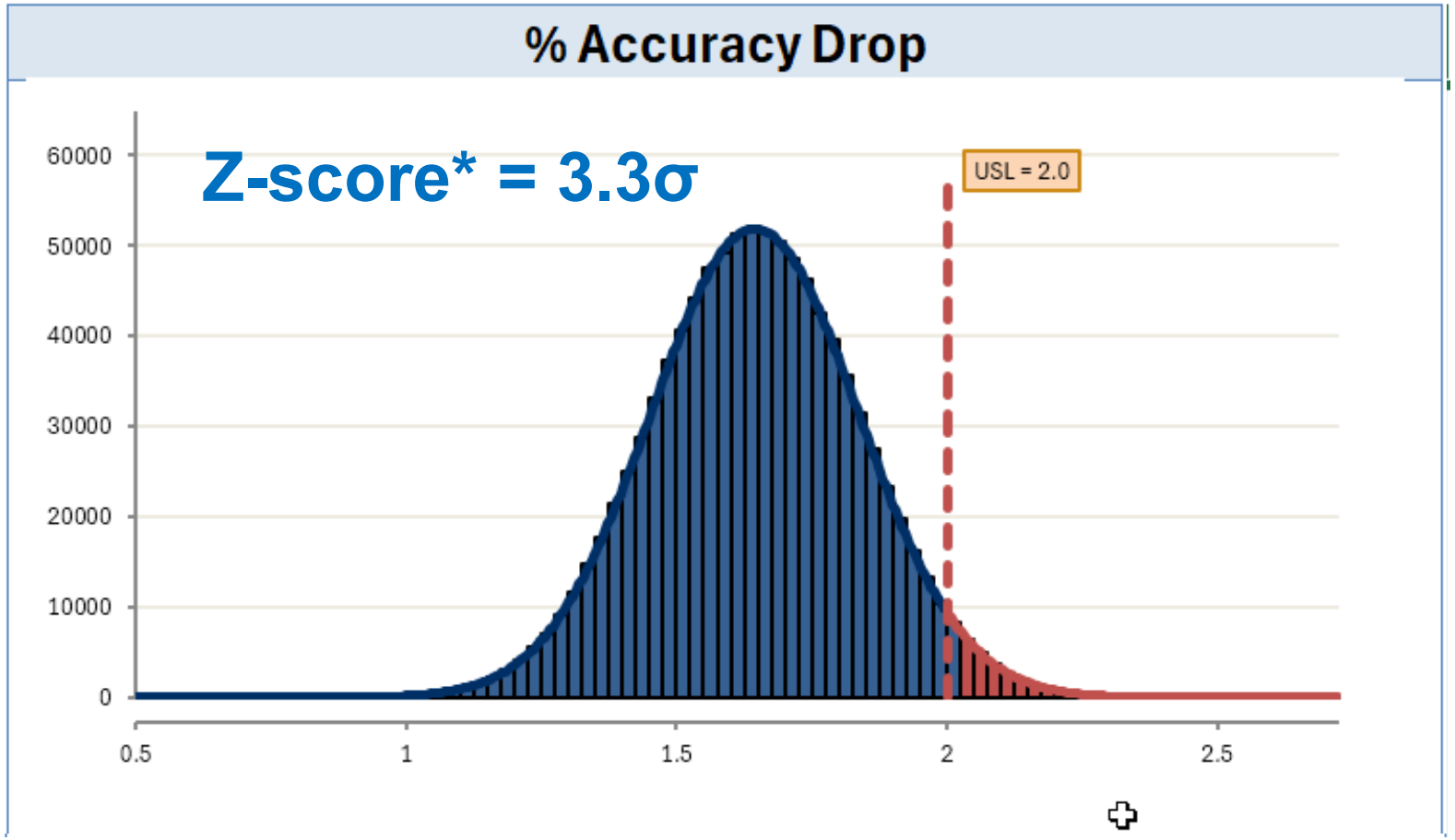
Normal Distro Statistics	
KS Test p-Value (Normal)	Not Avail
dpm	75,914
Cpk	0.478
Cp	

Observed Defect Statistics	
Simulations outside of spec	76,023

***note: based on Observed DPM**

Process Inputs		
Batch Size Rows	Mean	StDev
Distribution: **Normal**	12	0
Quantization	Mean	StDev
Distribution: **Normal**	-1	0
Precision Mode	Mean	StDev
Distribution: **Normal**	0	0

Level	Quantization	Precision Mode
−1	INT4	FP16
0	INT8	TF32
+1	FP16	FP32



Process Outputs	
# of Simulations	1,000,000
Mean	1.64
StDev	0.19
Median	1.64
LSL	
USL	2

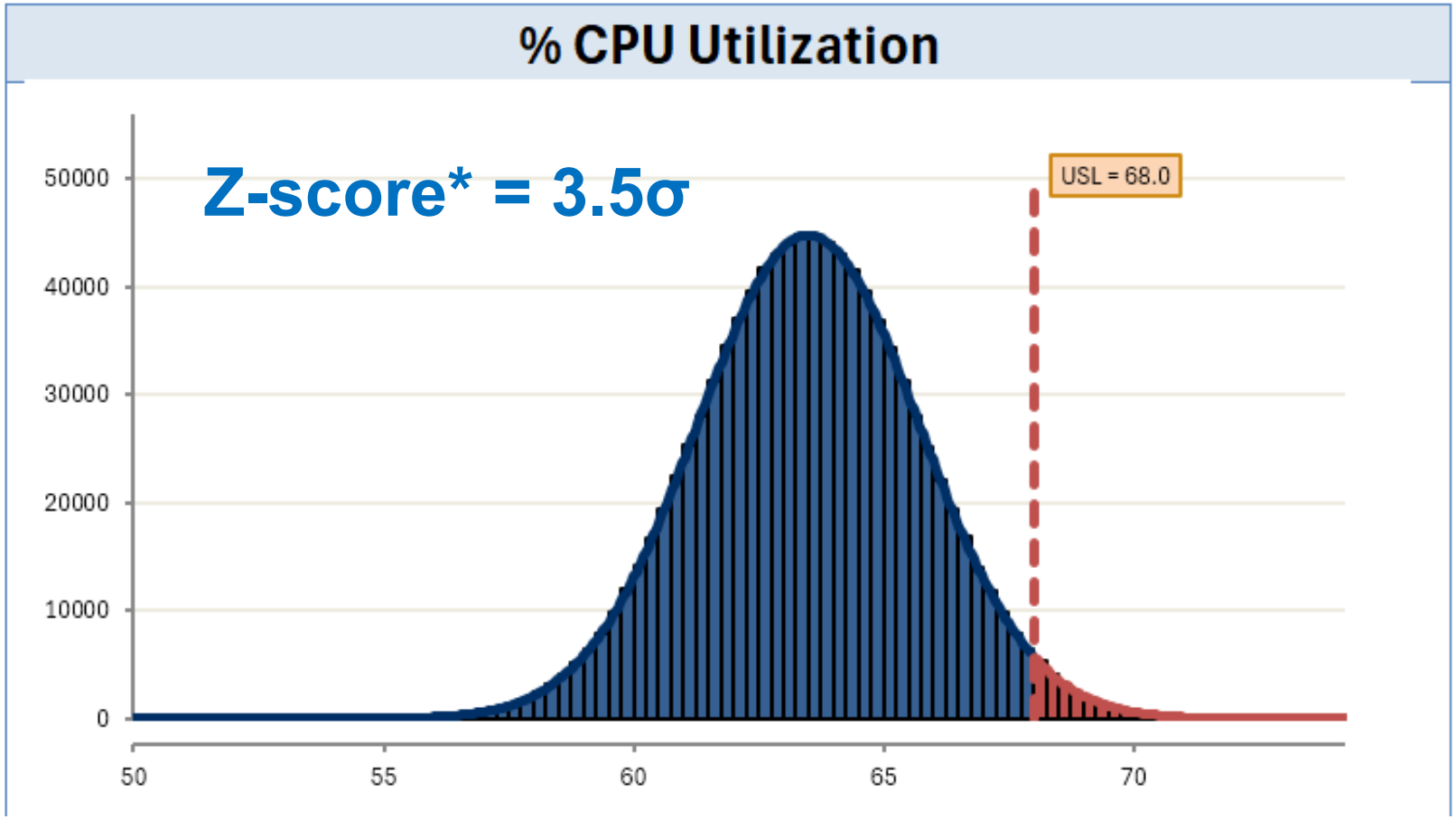
Normal Distro Statistics	
KS Test p-Value (Normal)	Not Avail
dpm	32,283
Cpk	0.616
Cp	

Observed Defect Statistics	
Simulations outside of spec	32,176

***note: based on Observed DPM**

Process Inputs		
Batch Size Rows	Mean	StDev
Distribution: **Normal**	12	0
Quantization	Mean	StDev
Distribution: **Normal**	-1	0
Precision Mode	Mean	StDev
Distribution: **Normal**	0	0

Level	Quantization	Precision Mode
−1	INT4	FP16
0	INT8	TF32
+1	FP16	FP32



Process Outputs	
# of Simulations	1,000,000
Mean	63.50
StDev	2.23
Median	63.50
LSL	
USL	68

Normal Distro Statistics	
KS Test p-Value (Normal)	Not Avail
dpm	21,546
Cpk	0.674
Cp	

Observed Defect Statistics	
Simulations outside of spec	21,829

*note: based on Observed DPM

Summary

OPTIMAL PARAMETERS

- Batch Size = 12 per inference call
- Quantization = INT4
- Precision Mode = TF32

CAPABILITY PER PERFORMANCE GOAL

- Inference Latency (USL=11ms): 3.9σ
- Throughput (LSL=125 ips): 2.9σ
- Accuracy Drop (USL= 2%): 3.3σ
- CPU Utilization (USL=68%): 3.5σ

Conclusion:

- **Multi-objective optimization** was realized by creating four transfer functions based on empirical data.
- The Monte Carlo simulation served two purposes:
  1. It modeled the residuals, effectively **achieving an R-squared value of 100%**.
  2. It facilitated capability analysis using **one million samples**.
- **Quantum XL Version 18** is a powerful tool for **ML Inference Tuning**, as it integrates design of experiments with Monte Carlo simulation.