

PVD Experiment Designer & Analyzer

v1.0, June 2025

This document provides an overview of the `pvD_DoE_v1.html` web-based tool. The tool is designed to assist researchers in designing, analyzing, and optimizing experiments based on Response Surface Methodology (RSM) for Physical Vapor Deposition (PVD) processes.

1. Overview & Workflow

The application guides the user through a structured workflow for performing a Design of Experiments (DoE) analysis for a PVD process. The primary goal is to find the optimal set of process parameters (factors) to maximize one or more desired outcomes (responses), such as film thickness, purity, and conductivity.

The user workflow is as follows:

1. **Design Experiment:** The user defines the process factors (e.g., Temperature, Pressure), their minimum and maximum operating values, and selects a DoE method. The tool provides default factors for Pressure, Substrate Temp, O₂/Ar Ratio, Power, Source Distance, and Deposition Time.
2. **Generate Plan:** The tool generates a detailed experimental plan, listing the specific combination of factor settings for each run.
3. **Enter Data:** The user performs the experiments and enters the measured responses (Thickness, Purity, and Conductivity) into the data table. A feature to fill the table with simulated data based on a pre-defined mathematical model is also available for validation and demonstration purposes.
4. **Analyze & Optimize:** The tool automatically performs a regression analysis to create a mathematical model linking the factors to the responses. It calculates key statistical metrics, predicts optimal conditions, and visualizes the relationships using contour plots.
5. **Export:** A summary report of the data, models, and predicted optima can be exported as a PDF document.

2. Design of Experiments (DoE) Methods

The tool implements three standard Response Surface Methodology (RSM) designs. RSM is a collection of statistical and mathematical techniques useful for developing, improving, and optimizing processes.

2.1. Full Factorial Design

A full factorial experiment measures the response at all possible combinations of the factor levels. For k factors, each at 2 levels (min and max), this design consists of 2^k runs.

- **Purpose:** Excellent for identifying the main effects of factors and their interactions.
- **Implementation:** The `generateFactorial` function creates this plan. It is primarily suited for fitting first-order models with interaction terms and serves as the foundation for the Central Composite Design.

2.2. Box-Behnken Design (BBD)

A Box-Behnken design is an efficient, three-level design used for fitting second-order (quadratic) models. It does not contain runs at the extreme vertices of the experimental space, which can be advantageous when these points represent expensive or unsafe operating conditions.

- **Purpose:** To fit a quadratic model and find optima without needing to test extreme factor combinations.
- **Structure:** The design is constructed by combining two-level factorial designs with incomplete block designs. The `generateBoxBehnken` function implements this by creating runs where factors are at their center points while pairs of other factors are at their min/max levels.
- **Requirement:** Requires at least 3 varying factors.

2.3. Central Composite Design (CCD)

A CCD is the most common design used for building second-order models. It is highly efficient and flexible.

- **Purpose:** To efficiently estimate a second-order model, allowing for the detection of curvature in the response surface and the location of optimal process settings.
- **Structure:** The `generateCentralComposite` function builds the design from three types of points:
 1. **Factorial Points:** The corners of the experimental space (coded as -1 and +1).
 2. **Center Points:** Replicate runs at the center of the domain (coded as 0).
 3. **Axial (or Star) Points:** Points along the axes of the factors, set at a distance ' α ' from the center, which enable the estimation of quadratic terms.
- **CCD Type (α value):**
 - **Face-Centered ($\alpha=1.0$):** The axial points are located on the “faces” of the factorial cube, confining all runs within the specified min/max range.
 - **Rotatable ($\alpha = k^{1/4}$):** Here, α is calculated from the number of factors (k) to ensure the prediction variance is constant at points equidistant from the center. This can result in axial points outside the min/max range.

3. Mathematical Foundation

3.1. Coding of Variables

To simplify calculations, real-world factor values (e.g., 400°C to 800°C) are “coded” into a dimensionless range from -1 to +1.

$$\text{Coded Value} = \frac{2 \times (\text{Real Value} - \text{Min Value})}{(\text{Max Value} - \text{Min Value})} - 1$$

The `uncodeValue` function reverses this process.

3.2. The Model

The application uses a second-order polynomial model. For two factors, X_1 and X_2 , the model is:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \epsilon$$

Where y is the predicted response (e.g., Thickness), β_0 is the intercept, β_i are linear coefficients, β_{ij} are interaction coefficients, β_{ii} are quadratic coefficients, and ϵ is the error.

3.3. Model Fitting: Method of Least Squares

The coefficients (β) are determined using the method of least squares, solved via the matrix equation:

$$b = (X^T X)^{-1} X^T y$$

Where \mathbf{b} is the vector of coefficients, \mathbf{X} is the design matrix, and \mathbf{y} is the vector of observed responses.

4. Analysis of Statistical Parameters

The `buildModel` function calculates several key statistics to evaluate the quality, significance, and predictive capability of the fitted second-order polynomial model.

4.1. R-Squared (R^2)

- **Definition:** Also known as the coefficient of determination, R-Squared measures the proportion of the total variation in the observed response variable that is explained by the model.
- **Formula Used:** $R^2 = 1 - \frac{SS_{Error}}{SS_{Total}}$, where SS_{Error} is the sum of squared residuals and SS_{Total} is the total sum of squares.
- **Interpretation:** The value ranges from 0 to 1 (or 0% to 100%). An R^2 of 0.90 means that 90% of the response's variability is accounted for by the model's factors. A higher R^2 is generally better, but it should not be the sole indicator of a model's quality.
- **Role:** Provides a quick assessment of the model's overall goodness-of-fit.

4.2. Adjusted R-Squared ($Adj R^2$)

- **Definition:** A modified version of R-Squared that is adjusted for the number of predictors (terms) in the model. It is more suitable for comparing models with different numbers of terms.

- **Formula Used:** $\text{Adj } R^2 = 1 - \left[\frac{(1-R^2)(n-1)}{n-p-1} \right]$, where n is the number of experimental runs and p is the number of model terms.
- **Interpretation:** The Adjusted R^2 increases only if the new term improves the model more than would be expected by chance. It is always lower than R^2 and is considered more reliable for evaluating model fit.
- **Role:** Helps prevent “overfitting” by penalizing the inclusion of unnecessary terms.

4.3. Predicted R-Squared (Pred R^2)

- **Definition:** This statistic measures how well the model is expected to predict responses for new observations that were not used in the model fitting.
- **Formula Used:** $\text{Pred } R^2 = 1 - \frac{\text{PRESS}}{SS_{\text{Total}}}$, where PRESS is the Predicted Residual Sum of Squares.
- **Interpretation:** A high Predicted R^2 indicates good predictive power. It should be in “reasonable agreement” with the Adjusted R^2 . A large difference (e.g., > 0.2) between them suggests the model may be overfit.
- **Role:** Provides the most rigorous test of the model’s predictive ability on new data.

4.4. PRESS (Predicted Residual Sum of Squares)

- **Definition:** A form of cross-validation used to calculate Predicted R^2 . It is calculated by summing the squares of the predicted residuals, where each residual is calculated from a model that omits that specific data point.
- **Interpretation:** A lower PRESS value is better, indicating smaller prediction errors. It is primarily an intermediate value for calculating Pred R^2 .
- **Role:** Serves as the foundation for assessing the model’s predictive performance.

4.5. Adequate Precision (Adeq. Precision)

- **Definition:** This statistic measures the signal-to-noise ratio. It compares the range of the predicted values at the design points to the average prediction error.
- **Formula Used:** $\text{Adeq. Precision} = \frac{\max(\hat{y}) - \min(\hat{y})}{\sqrt{\frac{p \cdot MS_{\text{Error}}}{n}}}$, where \hat{y} are the predicted values, p is the number of model terms, n is the number of runs, and MS_{Error} is the Mean Square Error.
- **Interpretation:** It indicates whether the model provides sufficient signal to be used for navigating the design space. A ratio greater than 4 is generally considered desirable. A value below 4 suggests the model has poor discriminative power.
- **Role:** A critical check to ensure the model is not just fitting noise.

4.6. Standard Deviation (Std. Dev.)

- **Definition:** The standard deviation of the residuals, representing the typical distance that the observed values fall from the fitted regression line.
- **Formula Used:** It is the square root of the Mean Square Error (MS_{Error}).
- **Interpretation:** A lower standard deviation is better, indicating a closer fit of the model to the data. This value is in the same units as the response variable.

- **Role:** Provides an absolute measure of the model's typical error.

4.7. Coefficient of Variation (C.V. %)

- **Definition:** The standard deviation expressed as a percentage of the mean of the observed responses.
- **Formula Used:** $C.V. \% = \frac{\text{Std. Dev.}}{\text{Mean}} \times 100.$
- **Interpretation:** It measures the relative amount of variation. A lower C.V. % is better. A value below 10-15% is often considered to indicate good model reproducibility.
- **Role:** Provides a standardized, relative measure of model error, independent of the response units.

4.8. Coefficient Statistics (Std. Error & 95% CI)

- **Definition:** These statistics assess the significance of each individual term in the model.
- **Formulas Used:**
 - **Standard Error:** The standard deviation of a coefficient's estimate, calculated as $\sqrt{MS_{Error} \cdot c_{ii}}$, where c_{ii} is the diagonal element of the $(X^T X)^{-1}$ matrix.
 - **95% Confidence Interval (CI):** The range within which the true value of the coefficient is likely to fall. It is calculated as Coefficient $\pm (1.96 \times \text{Std. Error})$. The tool uses a fixed t-critical value of 1.96 as an approximation.
- **Interpretation:** If the 95% CI range for a term contains zero, the term is generally not statistically significant. This means its effect cannot be distinguished from random noise.
- **Role:** Allows for model reduction by identifying and potentially removing non-significant terms.

5. Optimization

5.1. Single-Response Optimization

The `findGlobalOptimum` function seeks the combination of factor settings that maximizes a single response (Thickness, Purity, or Conductivity) using a random walk numerical optimization algorithm.

5.2. Multi-Response Optimization (Weighted Score)

The tool optimizes multiple responses using a weighted sum approach. The user assigns an importance weight to each of the three responses using sliders. The `findCombinedOptimum` function then maximizes a combined score to find a "compromise" setting that provides the best overall outcome.

$$\text{Score} = (W_{\text{thickness}} \cdot N_{\text{thickness}}) + (W_{\text{purity}} \cdot N_{\text{purity}}) + (W_{\text{conductivity}} \cdot N_{\text{conductivity}})$$

Where W is the weight for each response and N is the normalized predicted value for that response.

6. Visualization

The tool generates 2D contour plots to help visualize the response surface. A star symbol (★) on the plot marks the location of the predicted mathematical optimum for that response. When creating a 2D plot for a design with more than two factors, the other factors are held constant at their center point (average) value.

7. About the File

- **Filename:** pvD_DoE_v1.html
- **Version:** 1.0
- **Date:** June 27, 2025
- **Author:** NitaD, Univ Paris-Saclay