Inverse Functional Modeling for Drug Dissolution Profiles: A Statistical Framework for Curve-Based Formulation Design

by

Zheran Wang

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
May 2, 2026

Keywords: Dissolution testing, Functional data analysis, FPCA, f2 similarity, formulation optimization

Approved by

Mark Carpenter, Professor of Department of Mathematics and Statistics
Peng Zeng, Associate Professor of Department of Mathematics and Statistics
Hans-Werner van Wyk, Associate Professor of Department of Mathematics and Statistics
Roberto Molinari, Assistant Professor of Department of Mathematics and Statistics

Abstract

This work establishes a statistically reliable framework for reverse engineering drug formulations from a target dissolution profile, formulated as an inverse functional problem. Dissolution profiles describe the rate at which a drug is released from a dosage form and serve as a primary *in vitro* surrogate for product performance. Although regulatory evaluation commonly relies on scalar summaries such as the f_2 factor, dissolution behavior is inherently functional and reflects structured relationships between formulation variables and release dynamics.

The reverse engineering problem is formulated as an inverse mapping from formulation predictors to curve-valued responses under sampling noise and model-form uncertainty. By representing dissolution profiles as continuous functions of time, the framework enables dimension reduction, regression modeling, and optimization based on the entire trajectory. Parametric models and functional approaches based on spline smoothing and functional principal component analysis are integrated within a unified curve-space optimization architecture, with a reference-anchored extension to stabilize inverse estimation.

Applications to extended-release datasets demonstrate the behavior of the framework under real experimental conditions. To generalize beyond specific datasets, the framework is further evaluated through simulation studies spanning multiple mechanistic generating processes, including controlled structural misspecification. A formulation recovery error metric is introduced to assess inverse identification accuracy alongside curve-level discrepancy measures.

The results characterize when parametric efficiency is achieved under structural alignment and when functional representations provide greater robustness under heterogeneity or model mismatch. Collectively, this work provides a regulatorily grounded and general framework for curve-based dissolution modeling and formulation design when the underlying release mechanism is uncertain.

Artificial Intelligence (AI) Use Disclosure Statement

In the preparation of this thesis / dissertation, the following Artificial Intelligence (AI) tools were used: SciSpace, ChatGPT, Copilot, Claude. These tools were used primarily for literature searching, code generation, debugging, and code review. The author acknowledges full responsibility for the intellectual content of this work and has ensured that all AI-assisted sections have been reviewed and revised for accuracy and appropriate academic style. All AI-generated content was reviewed and validated for relevance, appropriateness, and accuracy before incorporation into the final document to maintain scholarly integrity of this research.

Acknowledgments

I would like to express my deepest gratitude to my advisor, Dr. Mark Carpenter, whose guidance shaped every aspect of this work. The countless hours we spent together working through statistical frameworks, debating modeling choices, and finding ways to bridge statistical methodology and pharmaceutical practice — were the foundation of this dissertation. His patience, rigor, and genuine investment in my development as a researcher made this work possible in ways that go far beyond what can be expressed here.

I also extend my sincere thanks to my committee members for their thoughtful engagement and constructive feedback throughout this process. Their questions pushed me to think more carefully and argue more precisely, and this dissertation is stronger for it.

This dissertation required me to contribute meaningfully to two disciplines, and I was fortunate to have expert guidance in both. On the pharmaceutical side, I owe special thanks to Professor Dongkai Wang and his research team at Shenyang Pharmaceutical University. Their generosity in sharing domain knowledge, their technical expertise, and their practical perspective on formulation development grounded this work in real pharmaceutical science. Learning to think like a formulation scientist — to understand not just the statistics but the underlying chemistry and regulatory context — changed how I approached every problem in this research. I am also grateful to Tao Zhang from Shenyang Dashan Pharmaceutical Technology Co., Ltd. and Yu Liu from the Research Institute of Northeast Pharmaceutical Group for their support and willingness to share data and insights related to pharmaceutical development.

Finally, I am grateful to my colleagues and friends who offered encouragement and perspective throughout this journey.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Problem Statement and Thesis Goal

The central problem addressed in this thesis is the *reverse engineering of drug formulations from a target dissolution profile*. Given an approved reference product (Drug A) with a known dissolution behavior, the objective is to identify formulation and process settings that produce a generic formulation (Drug B) whose dissolution profile closely matches the reference profile.

Traditional pharmaceutical approaches to dissolution analysis often rely on parametric models [25] and summary similarity metrics such as the similarity factor f_2 [42]. While these methods are widely used for modeling dissolution behavior and assessing profile similarity, they are not designed to solve the inverse problem of identifying formulation variables that produce a target dissolution profile.

From a statistical perspective, this problem can be formulated as an inverse design problem involving curve-valued responses, where formulation variables determine dissolution behavior observed at discrete time points under measurement noise and experimental variability. In practice, success is defined by satisfying regulatory and statistical similarity criteria rather than exact curve equality [63].

To address this challenge, this dissertation develops a unified optimization framework within which both parametric and functional modeling pipelines are constructed and compared, integrating curve representation, statistical modeling, and optimization to enable direct search for formulations that match a target dissolution profile.

Within this framework, two classes of modeling pipelines are developed: parametric pipelines (e.g., Weibull, KP) and functional pipelines based on FDA (FPCA). Both are coupled with a common search-based optimization strategy, enabling direct identification of formulations based on curve-level similarity.

The goal of this thesis is to develop a *unified statistical pipeline for formulation reverse engineering* that accommodates both parametric and functional modeling approaches. Dissolution profiles are treated as curve-valued responses, allowing the entire profile to be modeled and analyzed rather than relying solely on discrete summary statistics.

Through simulation studies and real dissolution datasets, this work evaluates the effectiveness of both parametric and FDA-based pipelines, each embedded within the same optimization framework, as flexible and robust approaches for curve-driven formulation design and reverse engineering.

1.2 Why Reverse Engineering Matters

The reverse engineering problem described in Section 1.1 arises naturally in formulation development settings, where multiple formulations may achieve similar dissolution behavior but differ in composition and process conditions.

Formulation reverse engineering plays an important role in both generic drug development and post-approval product modification. Although an approved reference product has demonstrated safety and efficacy, alternative formulations are often required in practice. These needs may arise from patent expiration, manufacturing constraints, cost considerations, supply-chain changes, or efforts to improve manufacturing robustness [62].

In generic drug development, manufacturers are not required to replicate the exact composition of the reference product. Instead, regulatory agencies require evidence that the test formulation exhibits comparable performance. Dissolution profiles therefore serve as a key intermediate indicator of product similarity during early-stage development.

From a regulatory perspective, independent formulation development is permitted once relevant patent and market exclusivity protections have expired. Generic manufacturers are allowed to develop their own formulations and manufacturing processes, provided that the resulting product demonstrates comparable quality, safety, and performance relative to the reference product. As a result, the goal of formulation development is typically not exact duplication of the original composition, but rather the design of an alternative formulation that achieves equivalent therapeutic performance.

This flexibility creates a large formulation design space in which many potential compositions may exist, yet identifying formulations that reproduce the desired dissolution behavior remains challenging in practice and often requires iterative experimentation and empirical adjustment.

In practice, dissolution-based reformulation is conducted within a structured evaluation pipeline. Candidate formulations are screened through laboratory experiments and compared to the reference product using *in vitro* similarity criteria before proceeding to more costly studies. Understanding how similarity is evaluated within this development pipeline is therefore essential for constructing statistically grounded approaches to formulation reverse engineering.

1.3 Evaluation Pipeline: From *in vitro* to *in vivo*

Within the formulation development pipeline described above, *in vitro* dissolution testing serves as the primary screening and decision-making tool prior to *in vivo* evaluation.

Compared with *in vivo* studies, dissolution tests are faster, less expensive, and free from ethical constraints associated with animal and human testing. As a result, regulatory agencies require that dissolution similarity be demonstrated before *in vivo* bioequivalence studies are conducted.

Dissolution profiles provide time-dependent information about drug release behavior and are commonly used as surrogates for *in vivo* performance. The relevance of dissolution behavior to bioequivalence has been discussed in the literature, and *in vitro–in vivo* correlation frameworks have been proposed to link dissolution data to pharmacokinetic outcomes [63].

Figure 1.1: Example *in vitro* dissolution testing apparatus used to measure drug release profiles from oral dosage forms. The apparatus maintains controlled temperature, agitation, and sampling conditions to monitor drug dissolution over time.

During *in vitro* evaluation, dissolution profile comparison tests are used to assess whether a test product exhibits release behavior similar to that of the reference product. These decisions are often based on a small number of sampling time points and summary similarity metrics, despite the inherently functional nature of dissolution profiles as continuous release curves over time.

The dissolution profiles in Figure 1.2 [32, 28] illustrate the characteristic monotone release behavior observed in oral dosage forms. Each curve represents the cumulative percentage of drug released over time, typically exhibiting a smooth increasing trajectory that approaches a plateau.

Notably, although the curves correspond to different drugs and experimental conditions, they can still exhibit similar overall shapes. This illustrates that different formulations may produce similar dissolution profiles.

1.4 Existing Formulation Development Approaches and Their Gaps

Having established the central role of *in vitro* dissolution comparison in formulation evaluation, this section reviews how formulation redesign is currently approached in practice and highlights the limitations of existing pharmaceutical and statistical methodologies.

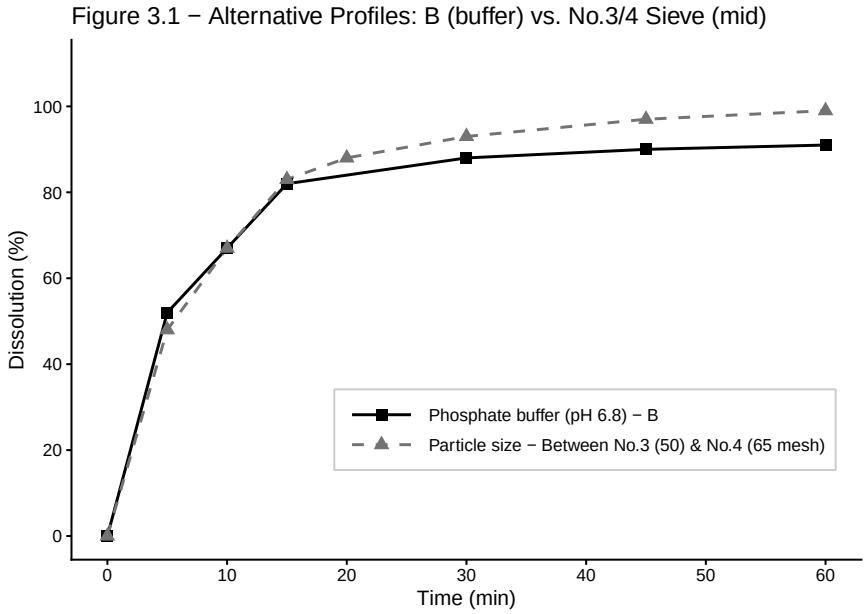


Figure 1.2: Example dissolution profiles illustrating the time-dependent release of drug from different formulations. Each curve represents the percentage of drug dissolved over time under standardized *in vitro* testing conditions.

1.4.1 Pharmaceutical development approaches

Current formulation development practices rely on a combination of pharmaceutical knowledge and empirical experimentation. Reference product labels provide information on excipients and dosage forms, allowing formulators to infer ingredient functionality. Patent literature may offer additional guidance on formulation and manufacturing processes.

When patents expire, reverse engineering techniques are often employed to characterize reference products. Quality by Design (QbD) [62] and Design of Experiments (DoE) [12] methodologies are then used to explore formulation and process variables in a structured manner.

While these approaches are effective in practice, they are often applied in an ad hoc or fragmented way and do not explicitly target curve-level similarity under uncertainty.

1.4.2 Statistical approaches used in practice

From a statistical standpoint, dissolution analysis and similarity assessment are typically addressed using separate approaches. For similarity evaluation, scalar summary metrics such as the f_2

similarity factor are widely used to quantify agreement between profiles [42]. Model-independent approaches, including distance-based measures and bootstrap-based extensions, have also been proposed when f_2 assumptions are violated.

In parallel, model-dependent approaches, such as the Weibull model, are used to represent dissolution behavior and analyze curve characteristics [25]. However, these models are generally applied for describing or fitting dissolution profiles rather than for formal similarity assessment or formulation optimization.

As a result, existing statistical approaches tend to address either profile comparison or curve modeling separately, without integrating these components within a unified workflow.

A more detailed review of statistical approaches to dissolution profile comparison is provided in Section 2.3.

1.4.3 Research gap

Despite extensive pharmaceutical and statistical research, there is no unified framework that treats dissolution profiles as functional objects while simultaneously supporting formulation redesign, uncertainty quantification, and regulatory decision-making. Existing methods either focus on empirical experimentation without formal statistical modeling or rely on scalar similarity metrics that discard important curve shape information.

This gap motivates the development of a unified statistical framework for the reverse engineering of drug formulations that integrates dissolution modeling, similarity evaluation, and formulation optimization within a single coherent workflow.

1.5 Thesis Contribution and Proposed Framework

This thesis develops a unified statistical framework for the reverse engineering of drug formulations based on dissolution profile analysis. The framework integrates functional data analysis, parametric modeling, and similarity-based evaluation within a single pipeline designed to support formulation redesign relative to a target reference profile.

Dissolution data are inherently functional, representing the evolution of drug release over time. The proposed framework treats the entire dissolution profile as the primary analytical object, enabling modeling, comparison, and optimization directly in curve space.

The proposed framework connects several components commonly used in pharmaceutical dissolution analysis—including Weibull modeling, similarity factor f_2, and functional data analysis (FDA)—within a systematic statistical workflow. By integrating these elements into a single pipeline, the framework enables coherent comparison of multiple modeling strategies and provides a principled approach to formulation redesign.

From a statistical perspective, the proposed framework can be viewed as comprising two interconnected components. The forward modeling component maps formulation variables to dissolution curves, while the inverse component seeks formulation settings that reproduce a target reference profile. Although these components are implemented within a unified pipeline, this forward–inverse distinction provides a useful conceptual framework for understanding the methodology.

While parametric models are effective when their structural assumptions are appropriate, their performance depends on correct model specification. In contrast, functional data analysis (FDA)-based approaches provide a flexible alternative that does not rely on a fixed parametric form.

This framework is designed to accommodate both modeling strategies, with particular emphasis on evaluating the robustness of functional representations under structural uncertainty.

The main contributions of this work are summarized as follows:

- **Problem formulation.** The thesis formulates the reverse engineering of drug formulations relative to a target dissolution profile as a statistical inverse design problem, in which formulation and process variables determine curve-valued responses.

- **Unified curve-based analysis framework.** A general pipeline is developed that represents dissolution profiles as functional objects, allowing curve-level modeling, comparison, and optimization while remaining compatible with regulatory similarity metrics such as the f_2 factor.

- **Comparative evaluation of modeling strategies.** Parametric approaches (such as Weibull-based models) and flexible functional representations (including spline smoothing and functional principal component analysis) are examined within a common analytical framework.

- **Optimization-based formulation identification.** A systematic strategy is introduced for identifying formulations whose predicted dissolution behavior closely matches a target reference profile. The approach combines functional regression, curve reconstruction, and similarity-based evaluation.

- **Application to experimental dissolution datasets.** The framework is applied to real extended-release dissolution datasets, demonstrating how the proposed pipeline can support formulation reverse engineering and complement existing industrial tools such as functional design of experiments.

- **Simulation-based robustness assessment.** Extensive simulation experiments are conducted to evaluate the behavior of the proposed pipelines under multiple dissolution mechanisms, sampling schemes, and noise structures, providing insight into robustness under model misspecification.

Together, these contributions establish a unified statistical pipeline for analyzing dissolution profiles and guiding formulation reverse engineering, while providing a systematic comparison between functional data analysis–based approaches and commonly used pharmaceutical modeling methods.

1.6 Thesis Organization

The remainder of this dissertation is organized as follows.

Chapter 2 reviews existing approaches for dissolution profile comparison, including regulatory similarity metrics, parametric modeling methods, and functional data analysis techniques relevant to pharmaceutical dissolution studies.

Chapter 3 introduces the proposed statistical framework for dissolution profile modeling and analysis. The chapter describes the curve-based representation of dissolution data, the modeling pipelines considered in this work, and the optimization strategy used to identify candidate formulations whose predicted dissolution profiles closely match a reference profile.

Chapter 4 applies the framework to experimental dissolution datasets, illustrating how the proposed methods perform in formulation analysis settings.

Chapter 5 presents simulation studies designed to evaluate the behavior and robustness of the proposed pipelines under controlled data-generating mechanisms representing several characteristic dissolution release patterns, as well as structural misspecification scenarios.

Chapter 6 discusses methodological insights, comparative performance across modeling approaches, and implications for dissolution-based formulation optimization.

Chapter 7 concludes with a summary of the main findings and outlines directions for future research.

Chapter 2

Literature Review

2.1 Chapter Roadmap

Dissolution profile comparison plays a central role in pharmaceutical development and regulatory decision-making. Over the past several decades, a range of statistical methods has been proposed to assess similarity between test and reference products, balancing simplicity, interpretability, and regulatory acceptance.

This chapter reviews the regulatory context and statistical foundations that motivate the methodology developed in this thesis. Emphasis is placed on model-independent similarity metrics, particularly the f_2 similarity factor, as well as extensions and critiques that highlight their strengths and limitations.

The review then expands to parametric and curve-based approaches, including functional data analysis methods that treat dissolution profiles as continuous objects rather than discrete measurements. These methods provide the conceptual foundation for the functional modeling, prediction, and optimization framework introduced in Chapter 3.

2.2 Regulatory Context for Dissolution and Similarity

In vitro dissolution testing plays a central role in pharmaceutical development by characterizing drug release behavior and serving as a surrogate for *in vivo* bioavailability, particularly in generic drug development. It is commonly used to evaluate whether a test formulation exhibits release characteristics comparable to an approved reference product and to ensure consistency across manufacturing batches [44].

Regulatory agencies establish dissolution specifications that must be met during the drug approval process [48]. For certain products, dissolution testing serves as a key component of regulatory evaluation, with agencies such as the U.S. Food and Drug Administration and the European Medicines Agency requiring evidence of similarity between test and reference profiles prior to *in vivo* bioequivalence studies [63, 10].

Despite dissolution profiles being inherently functional in nature, regulatory assessments are typically based on discrete-time summaries and scalar similarity metrics evaluated at matched sampling time points under standardized testing conditions.

2.2.1 The f_2 Similarity Factor: Use, Assumptions, and Limits

Among model-independent approaches, the f_1 difference factor and the f_2 similarity factor introduced by Moore and Flanner (1996) have received the most widespread regulatory acceptance, as they compare mean dissolution values directly without assuming an explicit parametric model. [42]

The f_2 similarity factor is defined as a logarithmic transformation of the sum of squared differences between mean dissolution values at matched time points. A value between 50 and 100 is generally interpreted as evidence of similarity between two dissolution profiles.

Due to its simplicity and ease of interpretation, the f_2 factor was formally adopted by regulatory agencies in 1997 as the primary tool for dissolution profile comparison. [63, 10]

However, the f_2 similarity factor is subject to several restrictive conditions that limit its applicability in practice. Regulatory guidance recommends that only one time point be included after both products exceed 85% dissolution, that variability be controlled such that no more than one time point exhibits a relative standard deviation greater than 10%, and that at least three time points be available for calculation [44, 41].

These requirements can be difficult to satisfy for drug products with high variability or complex release behavior, thereby reducing the reliability and applicability of the f_2 approach in such settings[16].

2.2.2 Extensions to f_2 and Alternative Similarity Frameworks

To address limitations of the f_2 factor, bootstrap-based methods have been proposed to incorporate uncertainty into similarity assessment. Bootstrap resampling has been used to construct confidence intervals for f_2, improving robustness under high variability [34, 40].

More recently, Liu et al. (2023) proposed a bias-corrected bootstrap f_2 statistic to reduce estimation bias and improve reliability. [29]

Tolerance interval methods provide an alternative reference-based framework for dissolution comparison. Zhai et al. (2016) proposed the use of tolerance limits derived from the reference batch to evaluate whether test profiles fall within statistically acceptable bounds. Building on this idea, Martinez and Zhao (2018) extended the approach by integrating optional f_2 comparisons with tolerance limits at individual time points, making the procedure particularly suitable for highly variable products [67, 35].

2.3 Statistical Approaches for Dissolution Testing

Methods for comparing dissolution profiles are generally classified into model-independent and model-dependent approaches. This distinction reflects whether an explicit mathematical model is assumed for the underlying dissolution process.

Comprehensive overviews of both model-independent and model-dependent approaches for dissolution profile comparison has been presented in the literature [21, 52].

2.3.1 Model-independent approaches

Model-independent approaches compare dissolution profiles directly using observed dissolution values at discrete time points, without assuming a specific functional form for the dissolution curve.

Early methods relied on pointwise statistical comparisons, such as analysis of variance (ANOVA), to assess differences between dissolution values at individual sampling times [36, 1]. However, such approaches do not provide a global measure of profile similarity.

Distance-based measures were subsequently proposed to summarize overall profile dissimilarity. Ma et al. (2000) introduced distance-based, model-independent similarity metrics, including the mean absolute difference and mean squared difference, as scalar measures of overall dissolution profile dissimilarity. [31]

Multivariate statistical methods treat dissolution measurements at multiple time points as correlated observations, enabling simultaneous analysis that accounts for the interdependence of the data. Classical techniques such as Hotelling's T^2 statistic have been explored in the literature as potential approaches for comparing dissolution profiles through tests on mean vector differences between formulations [53]. In addition, dimension-reduction approaches such as principal component analysis (PCA) have also been applied to summarize variability in dissolution data by treating measurements at discrete time points as multivariate observations [4]. However, such approaches operate on discretized observations and do not fully exploit the underlying functional structure of dissolution profiles.

Tsong, Sathe, and Shah (1996) introduced a multivariate framework for dissolution profile comparison based on the Mahalanobis distance, which accounts for correlation among measurements across multiple time points and provides a standardized measure of overall profile dissimilarity. This framework has since been adopted and extended in subsequent studies, demonstrating its usefulness for handling multivariate structure and variability in dissolution data [61, 19, 65].

A later review by Tsong et al. (2003) provided a comprehensive summary of distance-based approaches for dissolution profile comparison, including maximum and mean absolute differences, area-based distance measures, Rescigno indices and their weighted variants, as well as the standardized mean squared distance (Mahalanobis distance) proposed in earlier work [60].

More recently, Snee (2019) proposed a correlation-based approach for assessing dissolution profile similarity, in which each individual profile is compared against the batch mean to quantify

overall agreement. This approach provides a shape-sensitive alternative to traditional distance metrics. [58]

2.3.2 Model-dependent approaches

Model-dependent approaches assess dissolution similarity by fitting mathematical models to observed data and comparing either estimated parameters or reconstructed curves.

Parametric / mechanistic models Parametric models provide empirical representations of dissolution behavior using predefined functional forms. A wide range of models, including Weibull, zero-order, first-order, Korsmeyer–Peppas, and Hixson–Crowell formulations, have been widely used to capture different release patterns and kinetics. Due to their simplicity and interpretability, these models are commonly applied in dissolution analysis and formulation studies, where similarity is assessed through comparisons of fitted parameters or model-generated profiles [25, 55, 6, 39].

Mechanistic and kinetic-based models extend parametric approaches by incorporating assumptions about underlying drug release processes. These models aim to link dissolution behavior to physical mechanisms, providing additional interpretability at the cost of stronger modeling assumptions [6, 14].

More recently, stochastic modeling approaches have been proposed to capture variability and uncertainty in dissolution processes. These include probabilistic formulations based on Dirichlet distributions, gamma processes, and Wiener processes, offering flexible representations of dissolution behavior under random variation [49].

Regression / data-driven models In addition to parametric and mechanistic models, regression-based and data-driven approaches have been widely used to relate formulation variables to dissolution behavior. Linear and multivariate methods, such as multiple linear regression (MLR) and partial least squares (PLS), are commonly applied within designed experiments to model relationships between formulation factors and summary characteristics of dissolution profiles.

More flexible data-driven approaches, including machine learning methods such as artificial neural networks, have also been explored to model the mapping from formulation variables to dissolution responses. These methods provide strong predictive capability and can capture complex nonlinear relationships, but often require substantial data and may lack interpretability [43, 47, 15, 13].

In many cases, these approaches rely on intermediate parametric or low-dimensional representations of dissolution behavior, such as rate constants or fitted model parameters, rather than modeling the full functional profile directly [30].

While regression and data-driven methods provide flexible predictive frameworks, their performance may still depend on the validity of the underlying model representation and can be sensitive to sparse sampling and model misspecification.

2.3.3 Simulation-based frameworks for dissolution

Simulation-based approaches for dissolution profiles have been developed along several directions. Existing methods are typically based either on physics-driven mechanistic models, which simulate drug release using diffusion and mass-transport principles, or on empirical parametric models that approximate observed release behavior using predefined functional forms, such as Weibull or logistic-type models [64, 49, 29].

In addition, some studies incorporate formulation variables through designed experiments and data-driven modeling, where relationships between formulation factors and dissolution behavior are learned using linear and nonlinear regression techniques. In such settings, dissolution profiles are generated through model-based prediction, with curves reconstructed from fitted parameters or learned mappings, rather than simulated from a fully specified data-generating mechanism.

Across these approaches, generated curves are typically derived from assumed parametric structures or fitted models, and the underlying formulation–response relationship remains unknown. Consequently, these frameworks do not support controlled evaluation of model robustness, parametric misspecification, or inverse formulation recovery, as no ground truth is available for validation.

This limitation motivates the development of simulation-based frameworks in which the full data-generating mechanism is explicitly specified. Such frameworks enable systematic evaluation of modeling pipelines under controlled conditions, including assessment of robustness to model misspecification and the ability to recover underlying formulation variables.

2.4 Curve-Based and Functional Approaches in Dissolution

Dissolution profiles are inherently functional objects observed over time. Functional data analysis (FDA) provides a statistical framework for representing, smoothing, and modeling such data as continuous functions rather than discrete measurements. [50]

FDA methods developed in biological and longitudinal data settings have established rigorous theoretical foundations for smoothing, dimension reduction, and regression with sparsely observed curves. [5] These developments motivate the application of FDA to dissolution profiles, which exhibit similar characteristics, including sparse sampling, measurement noise, and between-unit variability.

By treating each dissolution profile as a realization of an underlying smooth function, FDA enables curve-level analysis that preserves global shape information and temporal structure. FDA techniques have been applied for curve smoothing, dimension reduction, regression modeling, and optimization in pharmaceutical and industrial contexts, providing a general statistical framework for analyzing complex process profiles. [20, 59]

FDA-based approaches support statistical operations that are difficult to perform reliably using discrete-time methods, including functional smoothing, dimension reduction, functional regression, and curve reconstruction. These tools make FDA particularly well suited for modeling dissolution behavior, where observations are often limited to a small number of time points. [7, 23, 27]

Functional principal component analysis (FPCA). Functional principal component analysis (FPCA) extends classical principal component analysis to functional data by representing variability among curves through a small number of orthogonal functional modes [50]. FPCA is particularly useful for dimension reduction when functional observations are sparsely sampled and subject to measurement noise. By projecting dissolution profiles onto a low-dimensional functional subspace defined by dominant modes of variation, FPCA provides a compact and interpretable representation of curve-level variability that is well suited for subsequent regression and formulation optimization tasks.

Recent advances in functional data analysis offer an alternative perspective by treating dissolution profiles as realizations of underlying smooth functions. In particular, FPCA has been shown to provide a robust and parsimonious representation of time-series response profiles, even in settings with sparse, noisy, or irregular sampling. McMullen et al. (2024) demonstrated that FPCA can accurately capture dominant modes of variation in complex pharmaceutical process profiles and, when combined with regression modeling, supports reliable prediction across multivariate design spaces [38]. In the context of dissolution modeling, Sousa (2025) further showed that FDA-based models incorporating FPCA achieved improved predictive performance compared to artificial neural networks, highlighting the advantages of function-level representations over pointwise prediction strategies [59].

Among the existing literature, the work of Kenett and Gotwalt (2023) is most closely related to the present dissertation. They compare functional data analysis and nonlinear regression for dissolution profile modeling and apply both approaches to the inverse formulation problem using the

same experimental dataset analyzed here. Using a Weibull growth model, they identify a formulation that reproduces a target dissolution profile and conclude that nonlinear regression outperforms FDA on this dataset [20]. However, their comparison is explicitly qualitative, with no quantitative evaluation metric reported. Moreover, the reported advantage of the Weibull model appears to be driven by a small subset of formulations near the optimal region, suggesting localized fit quality rather than consistent predictive performance across the formulation space. The authors further note that nonlinear regression depends on selecting an appropriate parametric form, and that in the absence of such prior knowledge, FDA-based approaches may be preferable. In addition, because the reference product corresponds to a proprietary commercial formulation, no ground truth is available to validate the recovered formulation, leaving the inverse problem unresolved. These limitations motivate the present work, which introduces a quantitative evaluation framework, conducts a systematic comparison across multiple modeling pipelines, and uses simulation studies to assess robustness under parametric misspecification and enable direct validation of inverse formulation recovery.

2.5 Summary and Research Gap

Existing dissolution comparison methods are primarily designed for regulatory decision-making and rely on scalar metrics or low-dimensional parametric summaries. While effective for pass/fail assessment, these approaches provide limited support for systematic modeling of dissolution behavior, particularly when full curve information and uncertainty are of interest.

More broadly, existing modeling approaches—including both parametric and functional methods—often focus on reproducing dissolution profiles under specific model assumptions or on predicting summary parameters from formulation variables. As highlighted by recent work (e.g., Kenett and Gotwalt (2023) [20]), comparisons between modeling approaches are frequently qualitative, rely on limited regions of the design space.

These limitations point to three key gaps:

- lack of quantitative and systematic evaluation of modeling performance,

- sensitivity to model misspecification and reliance on correct parametric forms,
- absence of ground truth in real data for validating inverse formulation results.

This gap motivates the development of a unified framework that combines functional data analysis with simulation-based evaluation. In this framework, dissolution profiles are treated as functional objects and linked directly to formulation variables, while simulation is used to generate controlled data for assessing model performance under varying conditions.

The methodology developed in Chapter 3 builds on this perspective by enabling curve-level modeling, systematic comparison of alternative modeling pipelines, and formulation search within a unified framework.

Chapter 3

Statistical Framework

3.1 Chapter Overview

This chapter presents the statistical pipelines developed to model dissolution profiles and recommend formulation settings that closely match a selected reference product. The central objective of this thesis is to evaluate whether Functional Principal Component Analysis (FPCA) provides a curve-agnostic framework capable of handling diverse dissolution shapes while still producing reasonable and interpretable formulation recommendations.

Two modeling families are considered:

1. Parametric pipelines (Weibull and Korsmeyer–Peppas models),
2. FDA-based pipelines.

For both families, a reference-anchored extension is introduced to improve alignment with a chosen reference profile. This chapter describes the structure and mechanics of each pipeline. Performance comparisons and simulation-based evaluation are presented in later chapters.

3.2 Notation and Data Structure

Let $i = 1, \ldots, N$ index formulations (batches) and $j = 1, \ldots, m$ index sampling time points. The observed dissolution profile for formulation i is denoted

$$y_i(t_j), \qquad 0 = t_1 < t_2 < \cdots < t_m, \tag{3.1}$$

with $y_i(t_j) \in [0, 100]$ representing percent dissolved.

Let $x_i = (x_{i1}, \ldots, x_{ip})^\top$ denote the formulation and/or process predictors associated with formulation i.

A selected reference formulation is denoted $y_{\text{ref}}(t_j)$, and the primary task is to identify a formulation vector $x^\star$ whose predicted curve is closest to the reference curve.

3.3 Distance Criteria and Decision Matrix

Pipeline recommendations are ranked using two complementary criteria.

3.3.1 Primary Criterion: ISD-Based Curve Distance

Define the integrated squared difference [50] between a candidate curve $\hat{y}(t; x)$ and the smoothed reference curve $\hat{y}_{\text{ref}}(t)$:

$$D_{\text{ISD}}(x) = \int_{\mathcal{T}} \left(\hat{y}(t; x) - \hat{y}_{\text{ref}}(t)\right)^2 dt. \tag{3.2}$$

Throughout this thesis, D_{ISD} is used as the primary optimization objective. For interpretability, results in plots and tables are reported in terms of D_{RISD}, defined as the square root of D_{ISD}.

3.3.2 Secondary Criterion: Similarity Factor f_2

The similarity factor f_2 is defined according to guidance from the U.S. Food and Drug Administration [63] as:

$$f_2 = 50 \log_{10} \left\{ \left[1 + \frac{1}{n} \sum_{j=1}^{n} (R_j - T_j)^2 \right]^{-1/2} \times 100 \right\}. \tag{3.3}$$

Here,

- R_j denotes the mean percent dissolved of the reference product at time point t_j,
- T_j denotes the mean percent dissolved of the test (or predicted) product at time point t_j,
- n is the number of common time points used in the comparison.

The statistic f_2 ranges from 0 to 100, with larger values indicating greater similarity between the two dissolution profiles. In general, $f_2 \geq 50$ is interpreted as evidence that the two profiles are similar.

Although the default ranking is ISD-first-then-f_2, the framework can be configured for f_2-first-then-ISD when regulatory alignment is prioritized.

3.4 Shared Pipeline Structure

All pipelines follow a common workflow:

1. Raw dissolution observations and formulation predictors
2. Curve representation (parametric or smoothed functional form)
3. Modeling layer linking formulation to curve representation, where full quadratic models with two-way interaction terms are primarily considered
4. Prediction of candidate curves
5. Optimization over feasible formulation space
6. Ranking using ISD and f_2

3.5 Parametric Pipelines

Parametric pipelines assume a pre-defined functional form for the dissolution curve and estimate formulation-specific parameters for that model.

3.5.1 Three-Parameter Weibull Pipeline

Model Representation. The three-parameter Weibull dissolution model [25] used in this thesis is

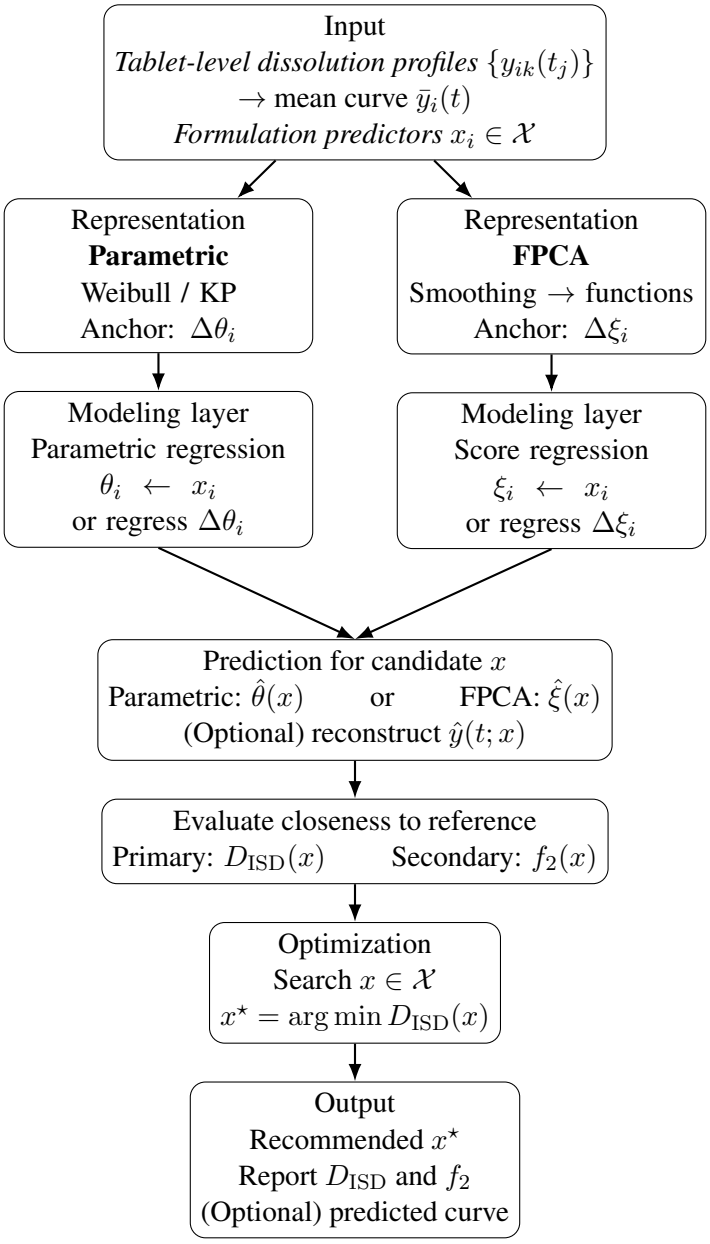


Figure 3.1: Unified workflow shared by parametric and FPCA pipelines, with optional reference anchoring incorporated within each representation path.

$$y_i(t) = a_i \left(1 - \exp\left[-\left(\frac{t}{b_i}\right)^{c_i}\right]\right), \tag{3.4}$$

where:

- $a_i > 0$ represents the asymptotic maximum release,

- $b_i > 0$ is a scale (time) parameter,
- $c_i > 0$ controls the curve shape.

This parameterization allows flexible concave and sigmoidal release patterns and corresponds directly to the implementation used in the computational pipeline.

Pipeline Structure.

1. Fit $(\hat{a}_i, \hat{b}_i, \hat{c}_i)$ for each formulation by nonlinear least squares [11]:

$$(\hat{a}_i, \hat{b}_i, \hat{c}_i) = \arg\min_{a,b,c} \sum_{j=1}^{m} \left[y_{ij} - a \left(1 - \exp\left\{ - \left(\frac{t_j}{b} \right)^c \right\} \right) \right]^2 .$$

2. Model the relationship between Weibull parameters and formulation predictors using transformed parameters (e.g., log-scale [26]):

$$\log(\hat{a}_i) = x_i^\top \beta_a + \varepsilon_{ia}, \tag{3.5}$$

$$\log(\hat{b}_i) = x_i^\top \beta_b + \varepsilon_{ib}, \tag{3.6}$$

$$\log(\hat{c}_i) = x_i^\top \beta_c + \varepsilon_{ic}. \tag{3.7}$$

3. For a candidate formulation x, predict the parameters via back-transformation[9] with log-normal bias correction:

$$\hat{a}(x) = \exp\big(x^\top \hat{\beta}_a + \hat{\sigma}_a^2/2\big), \tag{3.8}$$

$$\hat{b}(x) = \exp\big(x^\top \hat{\beta}_b + \hat{\sigma}_b^2/2\big), \tag{3.9}$$

$$\hat{c}(x) = \exp\big(x^\top \hat{\beta}_c + \hat{\sigma}_c^2/2\big), \tag{3.10}$$

where $\hat{\sigma}_a^2$, $\hat{\sigma}_b^2$, and $\hat{\sigma}_c^2$ denote the estimated residual variances from the corresponding log-linear regression models.

The predicted dissolution curve is then constructed as

$$\hat{y}(t;x) = \hat{a}(x)\left[1 - \exp\left\{-\left(\frac{t}{\hat{b}(x)}\right)^{\hat{c}(x)}\right\}\right].$$

4. Evaluate closeness to the reference using $D_{\mathrm{ISD}}(x)$ (primary) and $f_2(x)$ (secondary), and search over the feasible region $\mathcal{X}$ to obtain the recommended formulation $x^\star$.

3.5.2 Korsmeyer–Peppas (KP) Pipeline

Model Representation. The KP curve [22] used in this thesis is implemented as a capped power-law:

$$y_i(t) = \min\left(k_i\, t^{n_i},\, 100\right), \tag{3.11}$$

where $k_i > 0$ is a kinetic constant and $n_i > 0$ is the release exponent. The cap at 100 enforces the natural upper bound of percent dissolved.

Pipeline Structure.

1. Fit $(\hat{k}_i, \hat{n}_i)$ for each formulation by nonlinear least squares:

$$(\hat{k}_i, \hat{n}_i) = \arg\min_{k,n} \sum_{j=1}^{m} \left[y_{ij} - k\, t_j^n\right]^2.$$

2. Model the relationship between KP parameters and formulation predictors using transformed parameters (e.g., log-scale):

$$\begin{aligned} \log(\hat{k}_i) &= x_i^\top \beta_k + \varepsilon_{ik}, \\ \log(\hat{n}_i) &= x_i^\top \beta_n + \varepsilon_{in}. \end{aligned} \tag{3.12}$$

3. For a candidate formulation x, predict the parameters via back-transformation with log-normal bias correction:

$$\hat{k}(x) = \exp\big(x^\top \hat{\beta}_k + \hat{\sigma}_k^2/2\big), \tag{3.13}$$

$$\hat{n}(x) = \exp\big(x^\top \hat{\beta}_n + \hat{\sigma}_n^2/2\big), \tag{3.14}$$

where $\hat{\sigma}_k^2$ and $\hat{\sigma}_n^2$ denote the estimated residual variances from the corresponding log-linear models.

The predicted dissolution curve is then constructed as

$$\hat{y}(t; x) = \min\Big(\hat{k}(x)\, t^{\hat{n}(x)},\, 100\Big)\,.$$

4. Evaluate closeness to the reference using $D_{\mathrm{ISD}}(x)$ (primary) and $f_2(x)$ (secondary), and search over the feasible region $\mathcal{X}$ to obtain the recommended formulation $x^\star$.

3.5.3 Reference-Anchored Parametric Extension

Let $\hat{\theta}_{\mathrm{ref}}$ denote the fitted parameter vector of the reference profile (e.g., Weibull or KP parameters).

Define parameter differences:

$$\Delta\theta_i = \hat{\theta}_i - \hat{\theta}_{\mathrm{ref}}. \tag{3.15}$$

Instead of modeling $\hat{\theta}_i$ directly, the anchored pipeline models $\Delta\theta_i$ as a function of formulation predictors. Optimization then targets

$$\Delta\theta(x) \approx 0,$$

which enforces direct alignment to the reference curve in parameter space.

3.6 FPCA Pipelines

Functional principal component analysis (FPCA)[50] pipelines treat dissolution profiles as functional data without imposing a specific parametric form.

The FPCA pipeline proceeds through four sequential steps:

1. Smoothing discrete dissolution measurements to obtain continuous functional representations;
2. Functional decomposition via FPCA to obtain a low-dimensional set of scores (ξ_{ik}) that capture dominant modes of variation across curves;
3. Regression of FPCA scores on formulation variables;
4. Reconstruction of predicted dissolution curves for candidate formulations.

Each step is described in detail below.

3.6.1 Smoothing and Functional Representation

The first step converts discrete dissolution observations into continuous functional representations.

Each observed profile is smoothed to obtain a continuous function $\hat{y}_i(t)$ on a dense time grid. In this dissertation, smoothing is carried out using P-splines, which combine a cubic B-spline basis with a roughness penalty. Specifically,

$$\hat{y}_i(t) = \sum_{l=1}^{L} c_{il}\phi_l(t), \tag{3.16}$$

where $\{\phi_l(t)\}$ are cubic B-spline basis functions and c_{il} are coefficients estimated under a roughness penalty to control excessive local variation and enforce smoothness.

Alternative basis representations, including unpenalized B-splines and I-splines, were also explored and remain available within the proposed framework. Additional details on these alternative smoothing approaches are provided in Appendix A.

3.6.2 FPCA Decomposition

Given the smoothed functional data, the second step applies FPCA to identify dominant patterns of variation across curves.

Let $\mu(t)$ denote the mean function. The smoothed curves are decomposed as

$$\hat{y}_i(t) = \mu(t) + \sum_{k=1}^{K} \xi_{ik}\psi_k(t), \tag{3.17}$$

where:

- $\mu(t)$ is the mean function of the observed curves, representing the average dissolution profile across all samples;
- $\psi_k(t)$ are orthonormal eigenfunctions obtained from the covariance operator of the functional data, capturing the dominant modes of variation around the mean function;
- ξ_{ik} are the FPCA scores for observation i on the k-th component, determining how strongly the k-th eigenfunction $\psi_k(t)$ contributes to shaping the deviation of the individual curve from the mean function;
- K is the number of retained components, chosen as the smallest value such that the cumulative variance explained by the leading K principal components is at least 99.5%.

Here, FPCA is applied to the centered functions $\hat{y}_i(t) - \mu(t)$, which have zero mean by construction. Thus, the mean function $\mu(t)$ represents the overall average dissolution profile, while the FPCA components capture deviations around this mean.

Unlike the predefined spline basis functions used in smoothing, the eigenfunctions $\psi_k(t)$ are data-driven and adapt to the dominant modes of variation in the observed curve shapes. As a result, the FPCA representation does not rely on a predefined parametric form.

3.6.3 Score Regression and Curve Prediction

The third step links the functional representation to formulation variables by modeling the FPCA scores.

FPCA scores are linked to formulation predictors:

$$\xi_{ik} = x_i^\top \beta_k + \varepsilon_{ik}. \tag{3.18}$$

For a candidate formulation x, predicted scores $\hat{\xi}_k(x)$ reconstruct the full predicted curve:

$$\hat{y}(t; x) = \mu(t) + \sum_{k=1}^{K} \hat{\xi}_k(x)\psi_k(t). \tag{3.19}$$

This formulation enables prediction in the inverse problem setting, where candidate formulation vectors x are evaluated through their implied dissolution curves.

3.6.4 Reference-Anchored FPCA Extension

An extension of the above pipeline incorporates a reference curve to guide both decomposition and regression.

In the anchored FPCA pipeline, the reference curve is incorporated at two stages:

1. **FPCA Decomposition Stage:** The reference curve is included together with all other formulations when estimating $\mu(t)$ and $\{\psi_k(t)\}$. Therefore, the eigenfunctions and scores are influenced by the reference profile [57].

2. **Regression Stage:** Let $\xi_{\text{ref},k}$ denote the FPCA scores of the reference curve. Define

$$\Delta\xi_{ik} = \xi_{ik} - \xi_{\text{ref},k}. \tag{3.20}$$

The regression model is then fitted to $\Delta\xi_{ik}$ rather than ξ_{ik}.

Optimization targets

$$\Delta\xi(x) \approx 0,$$

which enforces alignment to the reference curve in functional score space.

By incorporating the reference curve both in the eigen-decomposition and in the regression modeling stage, the anchored FPCA pipeline shifts the optimization objective toward direct reference matching.

3.7 Optimization Strategy

All pipelines ultimately seek a formulation vector $x^\star$ that produces a dissolution curve closely matching the reference profile.

3.7.1 Inverse Formulation Problem

In both parametric and FPCA pipelines, the modeling layer defines a mapping from formulation space to a lower-dimensional representation:

$$x \longrightarrow \theta(x) \quad \text{or} \quad x \longrightarrow \xi(x),$$

where $\theta(x)$ represents parametric model parameters (e.g., a, b, c or k, n) and $\xi(x)$ represents FPCA scores.

Theoretically, one might attempt to solve the inverse problem directly:

$$\theta(x) = \theta_{\text{ref}} \quad \text{or} \quad \Delta\theta(x) = 0,$$

or

$$\xi(x) = \xi_{\text{ref}} \quad \text{or} \quad \Delta\xi(x) = 0.$$

However, the mapping from formulation space to parameter (or score) space is not necessarily one-to-one. In general, distinct formulation vectors x may produce identical or nearly identical $\theta(x)$ or $\xi(x)$ values.

3.7.2 Non-Uniqueness of the Inverse Mapping

As a consequence, solving the inverse equation

$$\theta(x) = \theta_{\text{ref}}$$

does not, in general, guarantee a unique solution for x. Even when a solution exists, multiple feasible formulations may satisfy the same parameter or score condition within the allowable formulation region $\mathcal{X}$.

This potential non-uniqueness reflects structural properties of the modeling framework, including:

- The representation of a functional dissolution profile using a finite set of model parameters or FPCA scores [8] (e.g., (a, b, c) or $\xi_1, \ldots, \xi_K$).,
- The possibility that the number of formulation predictors exceeds the dimensionality of the parameter or score space [17],
- Dependencies or correlations among formulation variables [3],
- Constraints defining the feasible region $\mathcal{X}$ [24].

3.7.3 Curve-Based Optimization Criterion

To resolve this ambiguity, optimization is performed in curve space rather than solely in parameter or score space.

For each candidate formulation x, the predicted curve $\hat{y}(t; x)$ is reconstructed and compared to the reference curve $y_{\text{ref}}(t)$ using the primary discrepancy measure

$$D_{\mathrm{ISD}}(x) = \int_{\mathcal{T}} \big(\hat{y}(t;x) - \hat{y}_{\mathrm{ref}}(t)\big)^2 \, dt$$

which represents the integrated area between the predicted and reference dissolution curves over the observed time grid $\mathcal{T}$.

It should be emphasized that D_{ISD} is a purely geometric distance measure defined in dissolution curve space. The notation reflects the integral form of the discrepancy rather than any pharmacokinetic interpretation.

The optimal formulation is then defined as

$$x^\star = \arg\min_{x \in \mathcal{X}} D_{\mathrm{ISD}}(x).$$

The similarity factor $f_2(x)$ is reported as a secondary metric.

By optimizing directly in curve space, the pipeline selects a formulation that is not only parameter-consistent but also minimizes the integrated difference from the reference profile over time.

Design Flexibility from Curve-Based Optimization. An additional advantage of optimizing directly in curve space is that the recommended formulation $x^\star$ need not coincide with the original reference formulation in predictor space. Because the mapping from formulation variables to dissolution behavior is not necessarily one-to-one, distinct formulation vectors may generate curves that are nearly indistinguishable with respect to D_{ISD} and f_2.

Consequently, the proposed framework can identify alternative formulation settings that achieve equivalent release behavior. From a pharmaceutical development perspective, this provides valuable flexibility: multiple candidate formulations may satisfy the same dissolution target while differing in excipient composition or processing conditions.

Rather than merely recovering an existing formulation, curve-based optimization expands the design space by revealing functionally equivalent but compositionally distinct solutions. This

property may support experimental redesign, robustness studies, and cost or manufacturability considerations.

3.7.4 Candidate Generation Strategy

Because the inverse mapping from formulation space to curve space is not analytically invertible, the optimal formulation $x^\star$ is obtained via search over a finite set of candidate points.

Candidate formulations are generated using Latin hypercube sampling (LHS) [37] over the feasible region $\mathcal{X}$ for the continuous predictors. For categorical variables, LHS is applied within each level, and the resulting samples are combined to form the full candidate set.

LHS provides stratified coverage in each formulation dimension while maintaining computational feasibility for large-scale search.

The same candidate set is evaluated within each pipeline for a given dataset or simulation replicate, and the formulation minimizing $D_{\mathrm{ISD}}(x)$ is selected as the recommended solution.

3.7.5 Potential Extension: Preference-Weighted Optimization

The optimization framework presented in this thesis focuses solely on curve-based similarity, with $D_{\mathrm{ISD}}(x)$ serving as the primary objective.

In formulation development, however, additional considerations may influence candidate selection. For example, a development team may prefer solutions near a target compression force due to manufacturing constraints, equipment calibration, or robustness considerations.

The proposed framework can be extended to incorporate such preferences by augmenting the ranking criterion with a predictor-space penalty term. A possible extension would define

$$J(x) = D_{\mathrm{ISD}}(x) + \lambda \sum_{j=1}^{p} w_j \big(x_j - x_{\mathrm{target},j}\big)^2, \tag{3.21}$$

where $\lambda \geq 0$ controls the trade-off between dissolution matching and proximity to preferred predictor settings, and $w_j \geq 0$ denotes a user-specified weight that reflects the relative importance or scaling of the j-th predictor. In practice, w_j can be used to prioritize certain formulation

variables (e.g., compression force) or to account for differences in units and variability across predictors.

This multi-objective formulation was not implemented in the current study but represents a natural direction for future development of the curve-based optimization framework.

3.8 Empirical Datasets

To evaluate the proposed pipelines under realistic extended-release conditions, two experimental dissolution datasets are analyzed in this study. Both datasets consist of multi-formulation extended-release (ER) products measured on common sampling grids.

3.8.1 Dataset A

Dataset A was obtained from Kenett and Gotwalt (2023) [20], who investigated extended-release dissolution behavior across multiple formulation batches. The dataset consists of 16 formulation batches, including one designated reference batch.

For each batch, dissolution was measured on six replicate tablets at a common sampling grid. Let $y_{ir}(t_j)$ denote the observed dissolution value for batch i and tablet replicate $r = 1, \ldots, 6$ at time t_j. The batch-level mean dissolution profile

$$\bar{y}_i(t_j) = \frac{1}{6} \sum_{r=1}^{6} y_{ir}(t_j)$$

is used as the input curve for all modeling and optimization procedures described in this thesis.

Each batch is characterized by four formulation or process predictors: the amounts of Polymer A and Polymer B, the total polymer content, and compression force. All predictors are standardized prior to regression modeling to ensure comparability of coefficient scales and numerical stability.

A designated reference batch is available in Dataset A and is treated as $y_{\mathrm{ref}}(t)$ for both the reference-anchored pipelines and the curve-based optimization procedure.

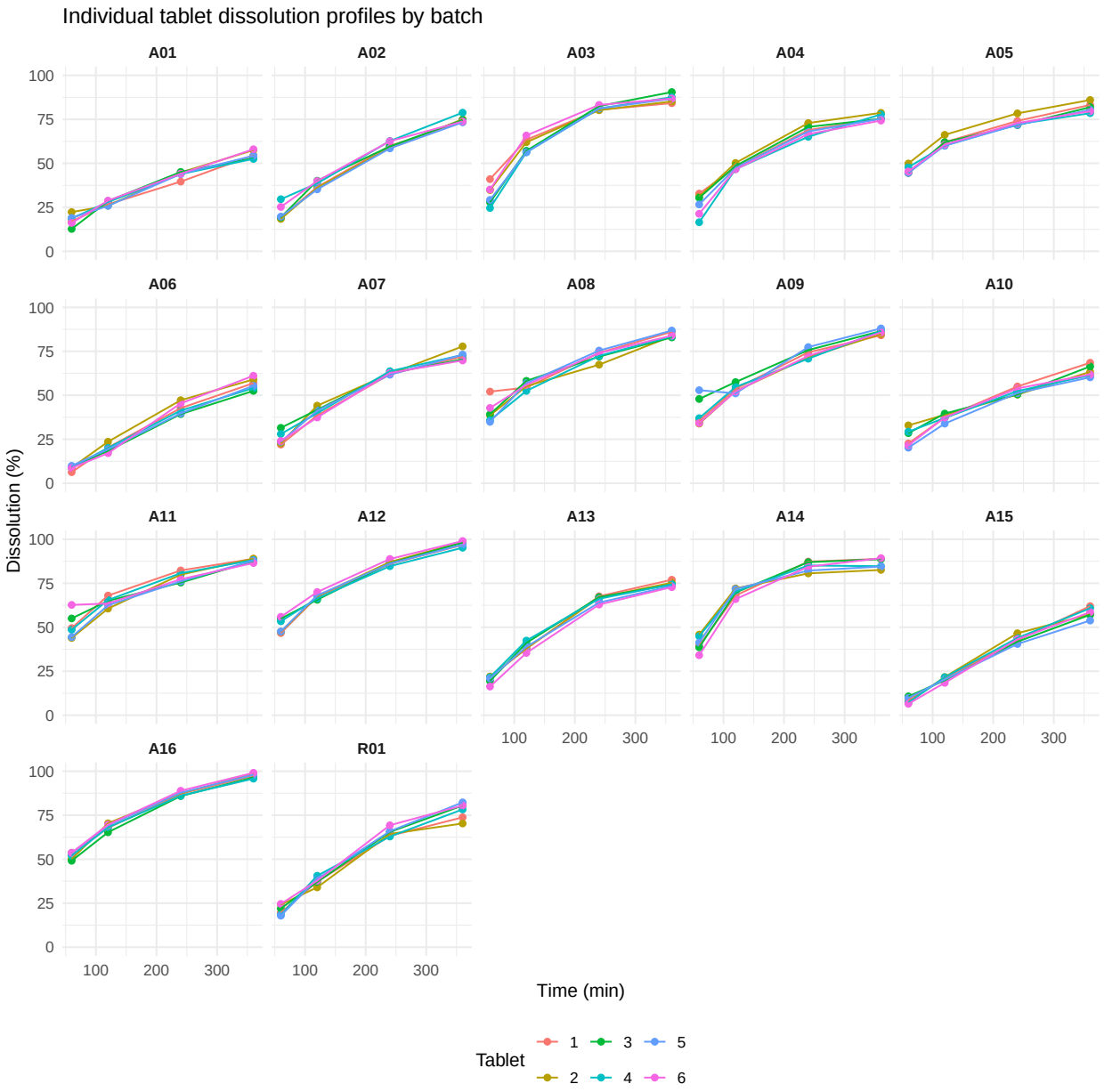


Figure 3.2: Example raw dissolution profiles from Dataset A. Replicate variability and shape diversity motivate flexible modeling.

3.8.2 Dataset B

Dataset B was obtained from Sousa et al. (2025) [59], in which the primary objective was to evaluate the ability of functional principal component analysis (FPCA) to accurately reproduce extended-release dissolution profiles.

Unlike Dataset A, Dataset B does not contain replicate tablet measurements per formulation. Each formulation is represented by a single observed dissolution curve measured on a common sampling grid. Let $y_i(t_j)$ denote the observed dissolution value for formulation i at time t_j. Because replicate data are not available, the observed curve is used directly as the input profile for modeling and optimization.

In addition, Dataset B does not include a designated reference product. For the purposes of reference-anchored analysis and optimization in this thesis, one formulation is selected as the reference curve,

$$y_{\text{ref}}(t),$$

and all remaining formulations are evaluated relative to this selected target profile.

This structural difference from Dataset A allows assessment of the proposed pipelines under a setting where (i) replicate averaging is not available and (ii) the reference curve must be defined analytically rather than supplied by experimental designation.

3.9 Chapter Summary

This chapter presented a unified statistical framework for curve-based dissolution modeling and formulation optimization. Parametric (Weibull and KP) and FDA-based pipelines were described, along with their reference-anchored extensions and a common candidate-search strategy based on Latin hypercube sampling.

The framework emphasizes optimization in curve space, allowing identification of functionally equivalent but potentially compositionally distinct formulations. Potential extensions, including preference-weighted multi-objective optimization, were also discussed.

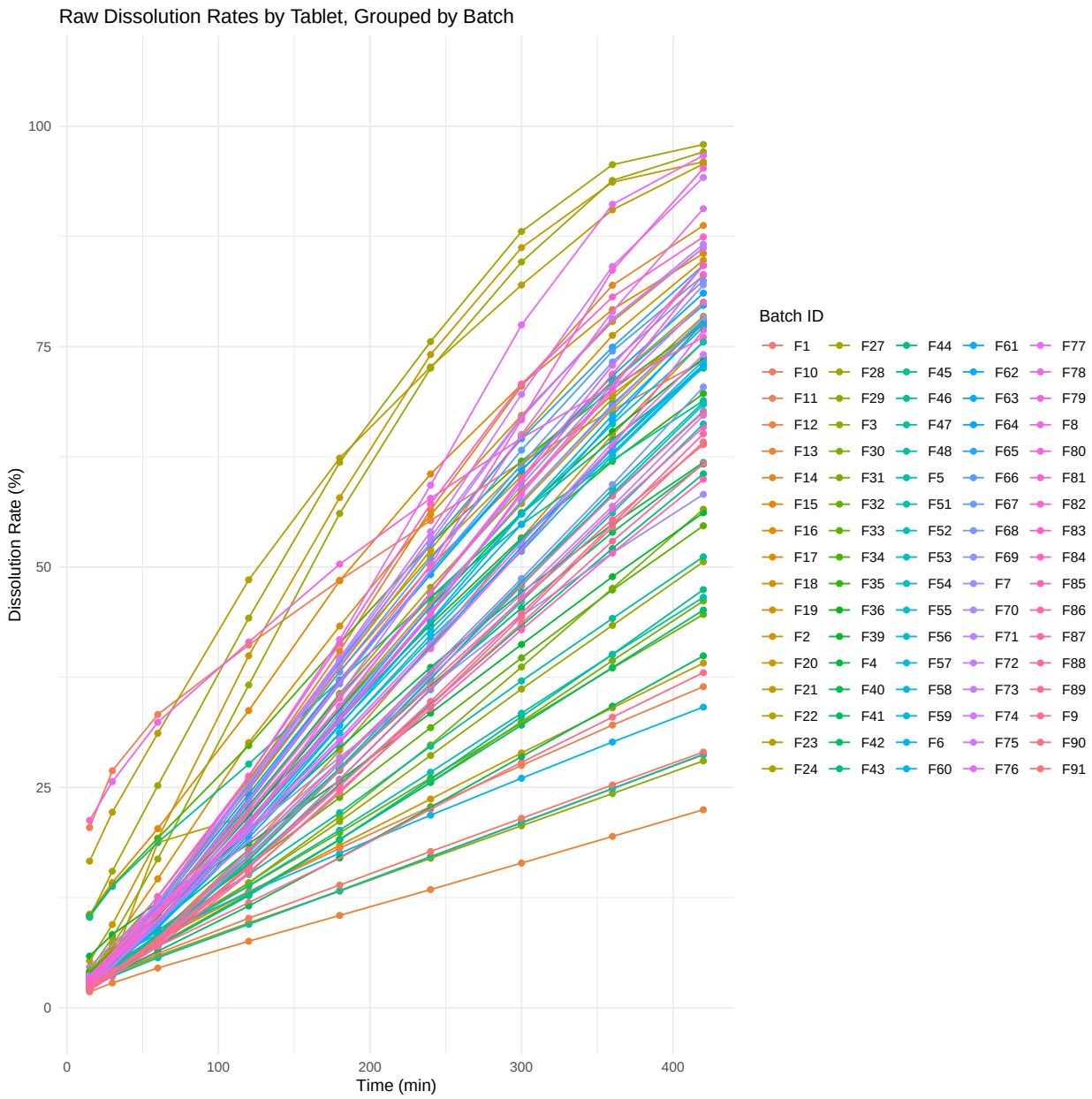


Figure 3.3: Example raw dissolution profiles from Dataset B, with one measurement per formulation, emphasizing modeling assumptions.

Chapter 4

Real Data Application

4.1 Overview of Real Datasets

This chapter applies the proposed dissolution modeling and optimization pipelines to two real extended-release datasets, denoted Dataset A and Dataset B.

Dataset A consists of multiple tablet replicates per formulation (batch), with six tablets available for each formulation. Batch-level mean dissolution profiles are computed and used for model fitting and optimization. Dataset A includes a reference formulation measured under the same conditions as the test formulations.

Dataset B contains a single observed dissolution profile per formulation, resulting in greater between-formulation variability and the absence of within-formulation averaging. No reference formulation is available in Dataset B.

These two datasets provide complementary evaluation settings for assessing the behavior of parametric and FDA-based pipelines under real-world conditions.

4.2 Analysis Framework for Real Data

All six pipelines considered in the simulation study are applied to each dataset:

- Weibull (Standard),
- Weibull (Reference-Anchored),
- KP (Standard),
- KP (Reference-Anchored),

- FPCA (Standard, P-spline smoothing),
- FPCA (Reference-Anchored).

The same optimization strategy described in Chapter 3 is employed. Curve similarity is evaluated using D_{RISD}, computed on a dense grid of 100 time points, along with the similarity factor f_2. Because the true formulation–curve relationship is unknown for real datasets, formulation recovery error cannot be directly assessed in this chapter.

4.3 Results for Dataset A

Dataset A, described in Section 3.8, is analyzed using the replicate-mean dissolution profiles for each formulation batch. A designated reference batch is available and is treated as $y_{\mathrm{ref}}(t)$ for anchoring and optimization.

4.3.1 Data Overview

4.3.2 Curve Fitting Comparison

Figure 4.2–4.4 present representative examples of fitted or reconstructed dissolution profiles from each modeling family, with the reference curve overlaid for comparison. To improve readability, two illustrative profiles are shown per modeling approach.

The full set of reconstruction results is deferred to Appendix B, Section B.1.1.

The comparison emphasizes curve representation quality, independent of downstream optimization performance.

The Weibull model captures the overall monotonic release behavior effectively, providing a smooth parametric representation across the observed time domain. Its three-parameter structure accommodates gradual curvature changes while maintaining interpretability.

The KP model provides reasonable approximation in the early and mid-release regions; however, minor deviations may occur at later time points due to its intrinsic power-law structure, which restricts flexibility in describing more complex curvature.

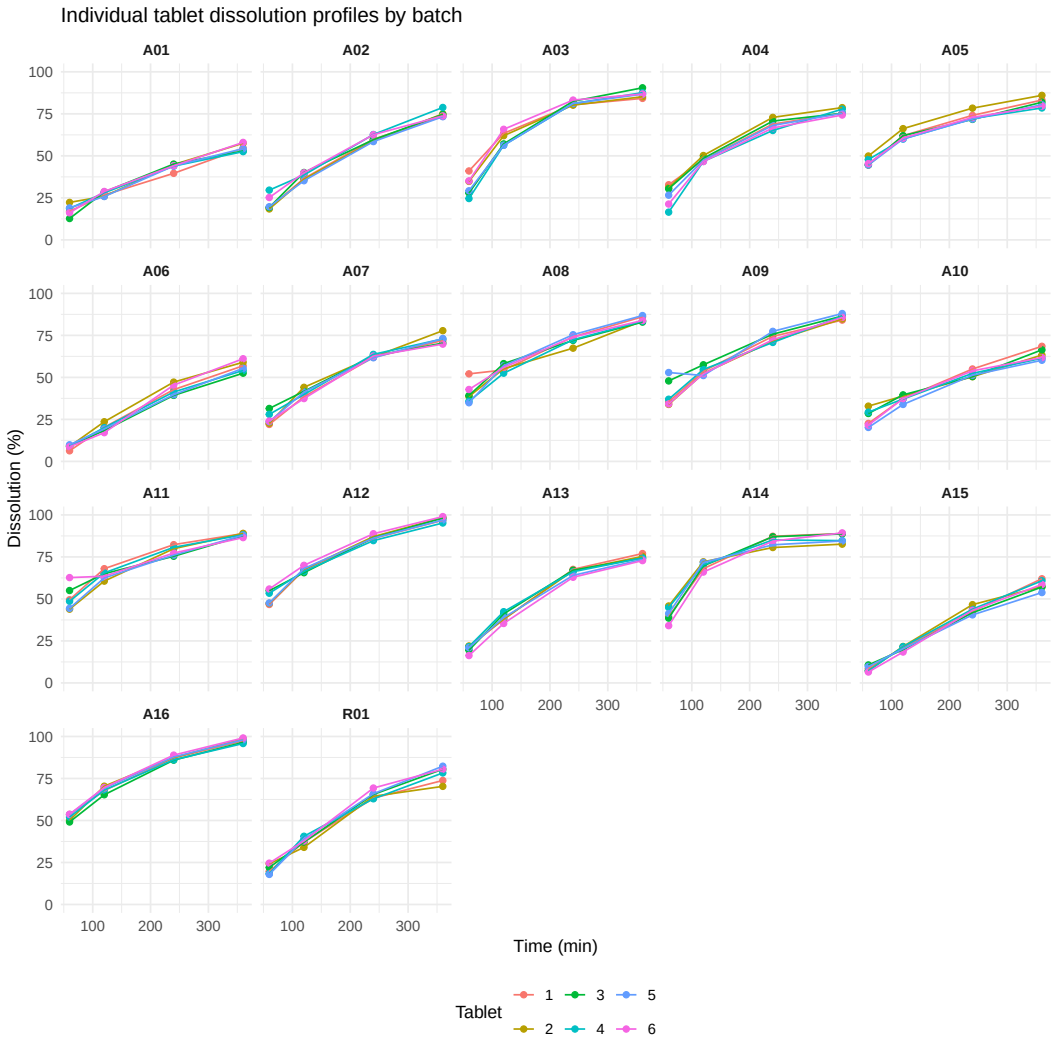


Figure 4.1: Raw dissolution profiles for Dataset A, where the reference curve is observed under the same experimental conditions as the candidate formulations. All formulations, including the reference, are measured with multiple tablet replicates, and batch-level mean curves are used for analysis.

The FPCA pipeline, implemented with P-spline smoothing, offers a fully data-driven reconstruction over the entire time domain. This approach adapts naturally to subtle curvature variations without imposing a predefined parametric form.

Although the anchored variants modify the regression target through reference-based differencing, both approaches are fitted using the same set of predictors in the full model. As a result, the fitted curves from the standard and anchored versions are often similar, and differences are not visually pronounced at the curve-fitting stage. However, when variable selection is introduced, the effect of anchoring becomes more evident, as the two approaches may select different subsets of predictors. These differences are further amplified in the optimization stage, where they lead to more distinct formulation recommendations and performance outcomes.

4.3.3 Optimization Outcomes

Optimization was conducted over the feasible formulation region $\mathcal{X}$ to identify

$$x^\star = \arg\min_{x \in \mathcal{X}} D_{\mathrm{RISD}}(x),$$

with f_2 reported as a secondary similarity measure.

In the candidate generation step, Latin hypercube sampling (LHS) was used to explore the continuous formulation space efficiently. However, LHS is designed for continuous variables and does not directly accommodate categorical factors. In Dataset B, the diluent variable is categorical. To incorporate this factor within the search procedure, diluent levels were randomly assigned to the candidate formulations generated from the continuous LHS design. With a large candidate set of 10,000 formulations, this random assignment provides adequate coverage of the categorical levels in practice. Alternatively, a stratified approach could be adopted in which separate LHS designs are generated within each diluent level and then combined. Both strategies ensure that the feasible formulation space is sufficiently explored while preserving the space-filling properties of the LHS design for the continuous variables.

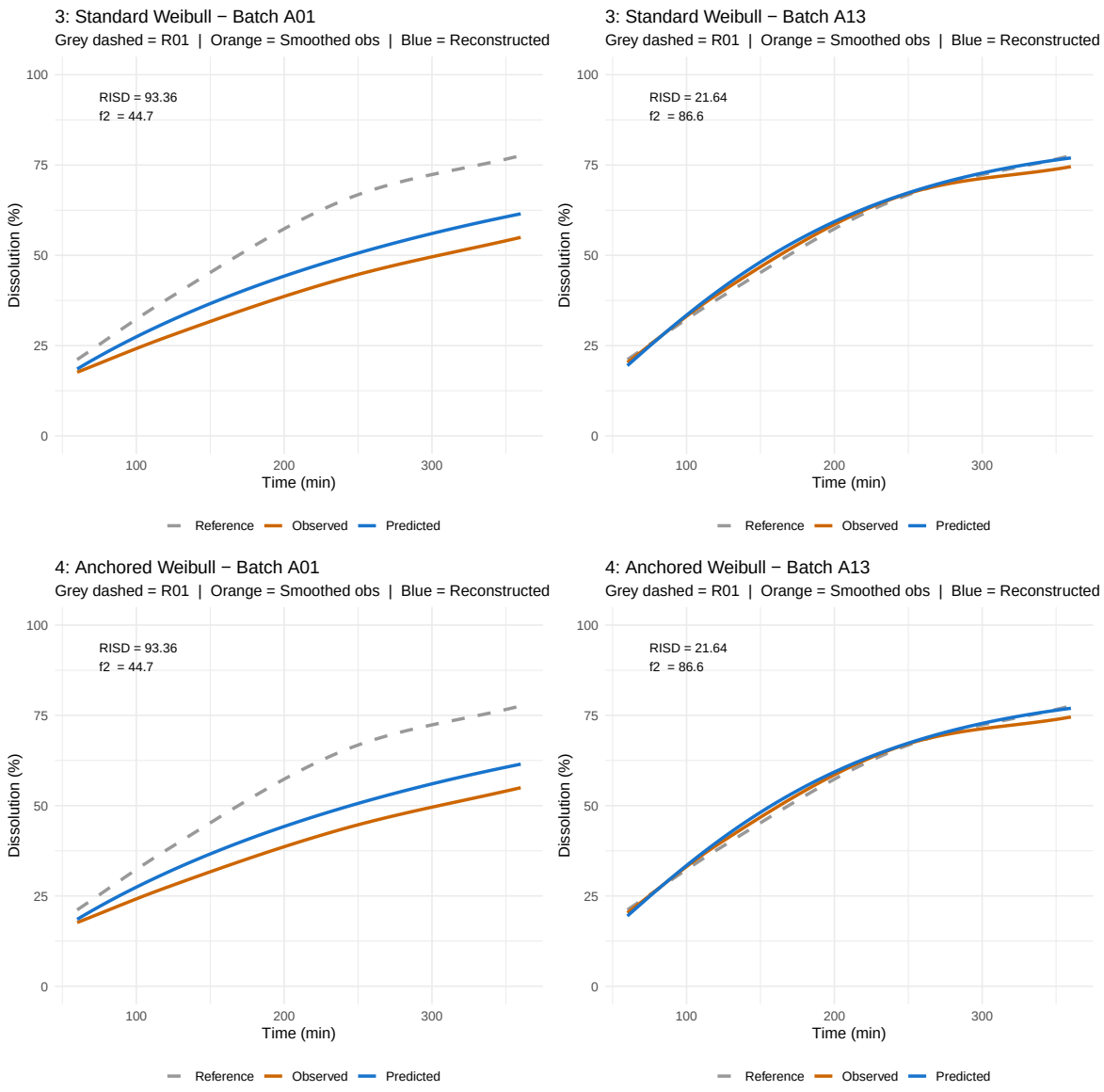


Figure 4.2: Dataset A: Fitted curves example under the Weibull model.

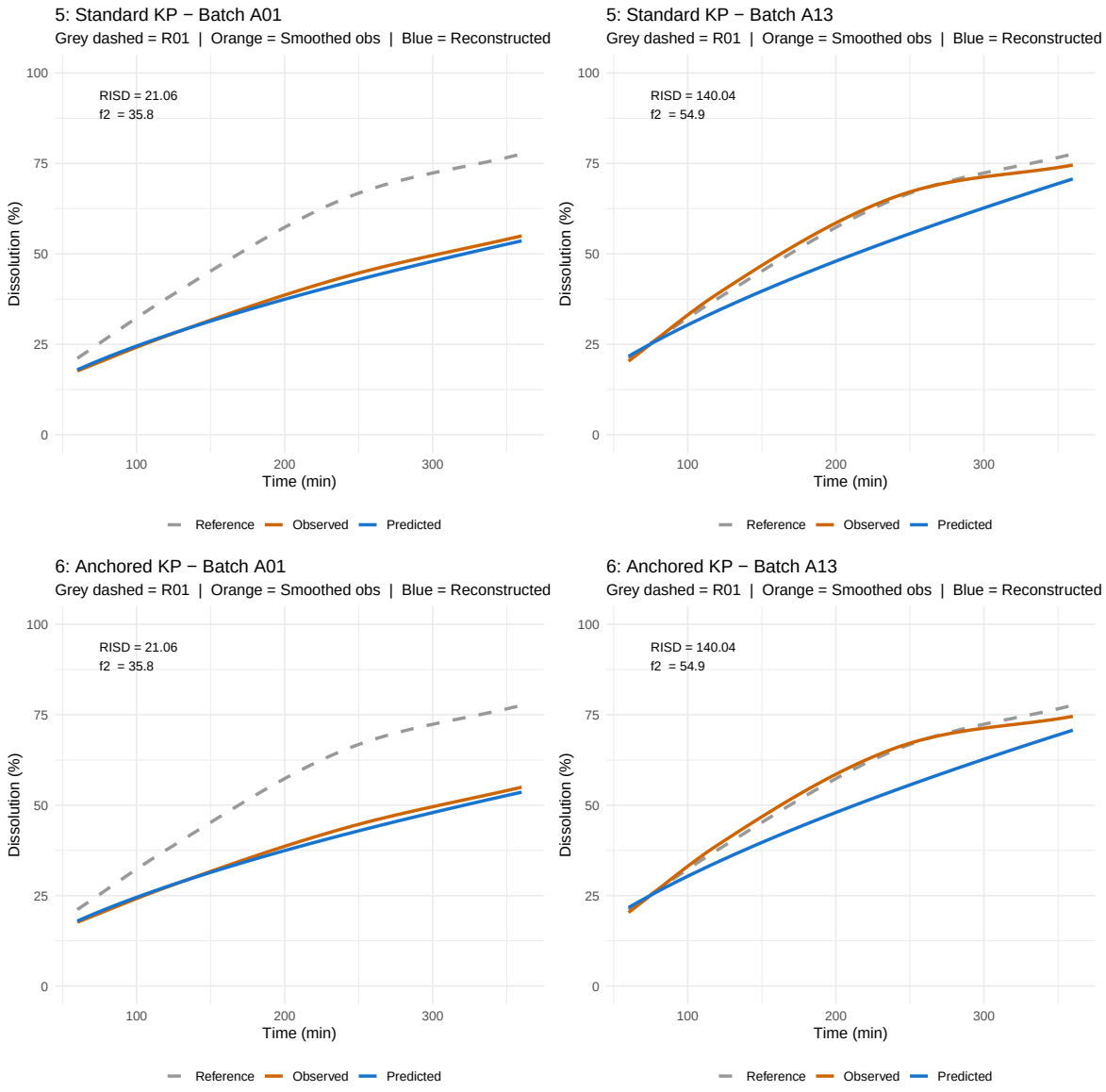


Figure 4.3: Dataset A: Fitted curves example under the KP model.

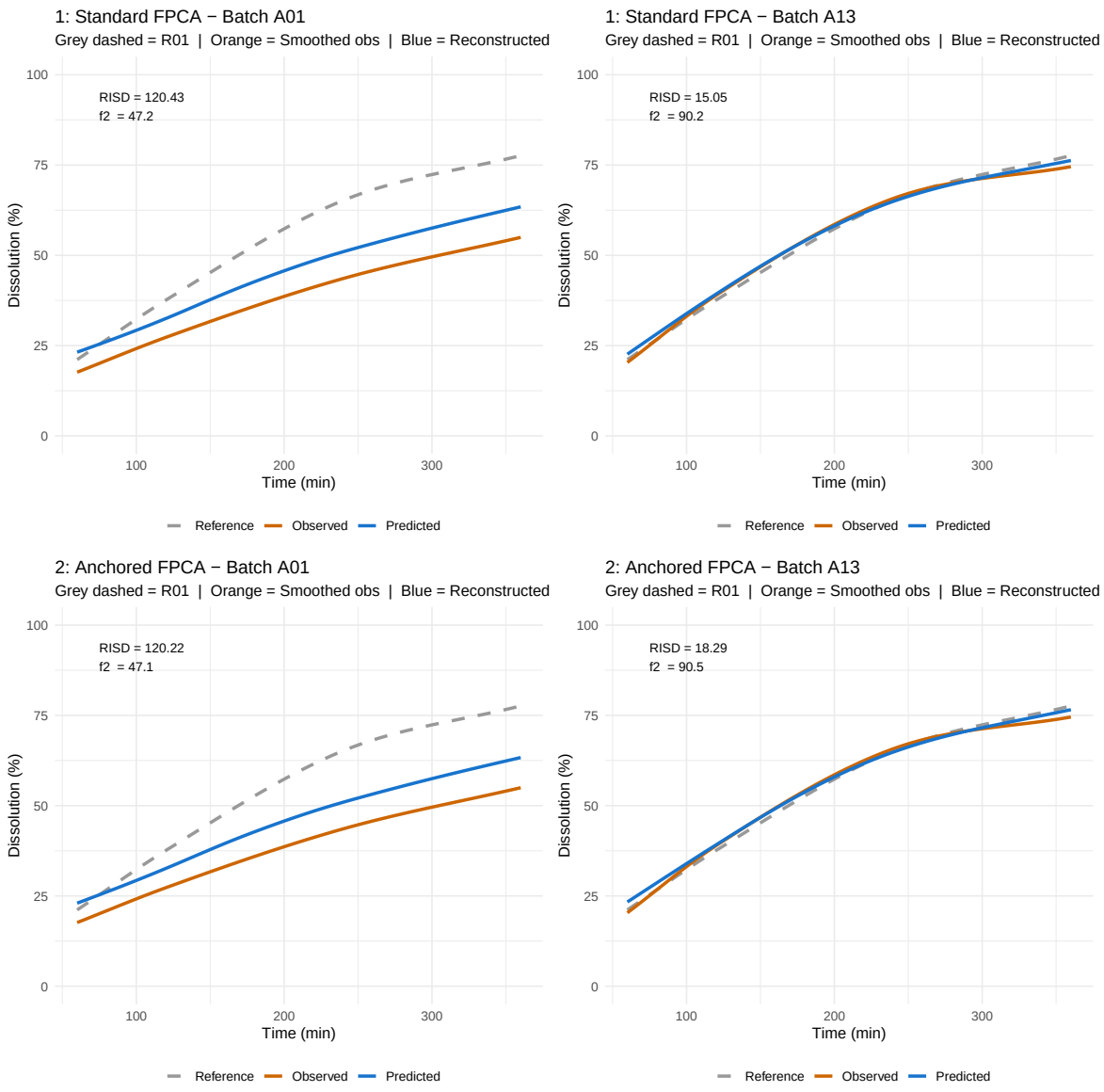


Figure 4.4: Dataset A: Fitted curves example under the FPCA (P-spline) model.

Figures 4.5–4.7 present the top-performing optimized candidate under each modeling pipeline, while Figure 4.8 summarizes the distribution of D_{RISD} values across the evaluated candidate set. Notably, the search-based framework identifies multiple near-optimal formulations with comparable performance, taking advantage of the non-uniqueness of the inverse mapping. These additional candidates are provided in Appendix B, Section B.2.1. Numerical summaries of the optimal formulations and their corresponding D_{RISD} and f_2 values are provided in Table 4.1. Table 4.2 presents the same results with an additional column indicating the variable selection method used for each model.

Across pipelines, the Weibull-based models achieved the lowest D_{RISD} values under both the full model and variable selection settings. In the full model, the standard and anchored Weibull pipelines produced identical results ($D_{\mathrm{RISD}} = 18.70$, $f_2 = 89.4$), indicating the closest agreement with the designated reference formulation. Under variable selection, the standard Weibull pipeline further improved to $D_{\mathrm{RISD}} = 13.13$ with $f_2 = 92.9$, maintaining the strongest overall performance.

The FPCA pipelines yielded moderately larger D_{RISD} values but still achieved strong similarity scores, with both standard and anchored variants satisfying the conventional f_2 criterion. In contrast, the KP-based pipelines exhibited substantially larger D_{RISD} values, indicating weaker optimization performance under this dataset.

Differences between standard and anchored variants are not observed under the full model, where both approaches yield identical results due to the underlying predictor structure. When variable selection is applied, however, anchoring produces modest improvements in some cases, particularly for the FPCA pipeline.

Overall, Dataset A demonstrates that the Weibull-based representation provides the best alignment with the reference profile, reflecting appropriate structural alignment between the model form and the observed dissolution behavior. The consistent performance across pipelines, together with the relatively small differences between standard and anchored variants, suggests that model adequacy plays a more prominent role than anchoring in determining optimization outcomes for this dataset.

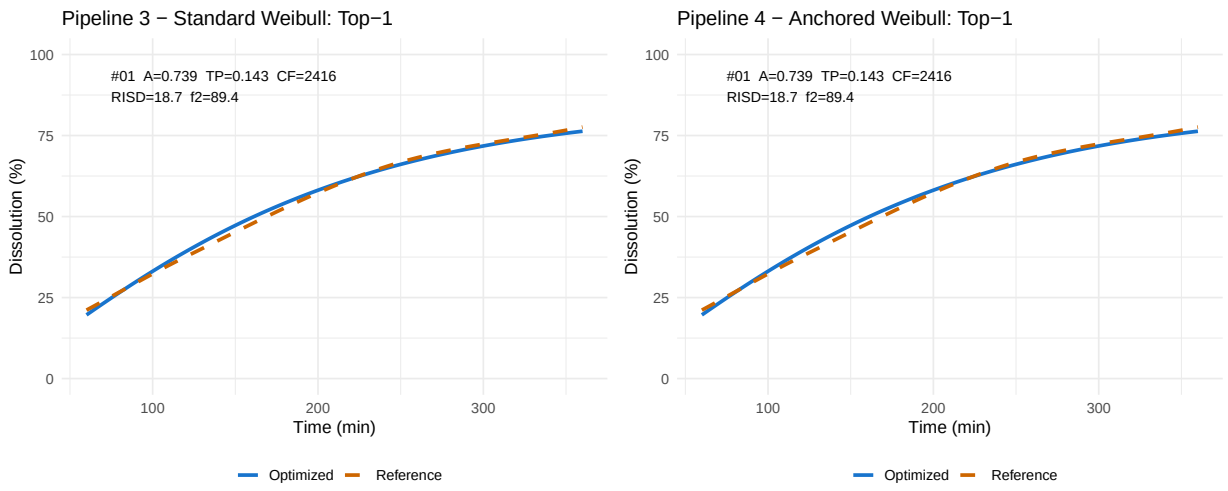


Figure 4.5: Dataset A: optimized dissolution curves under the Weibull pipeline.

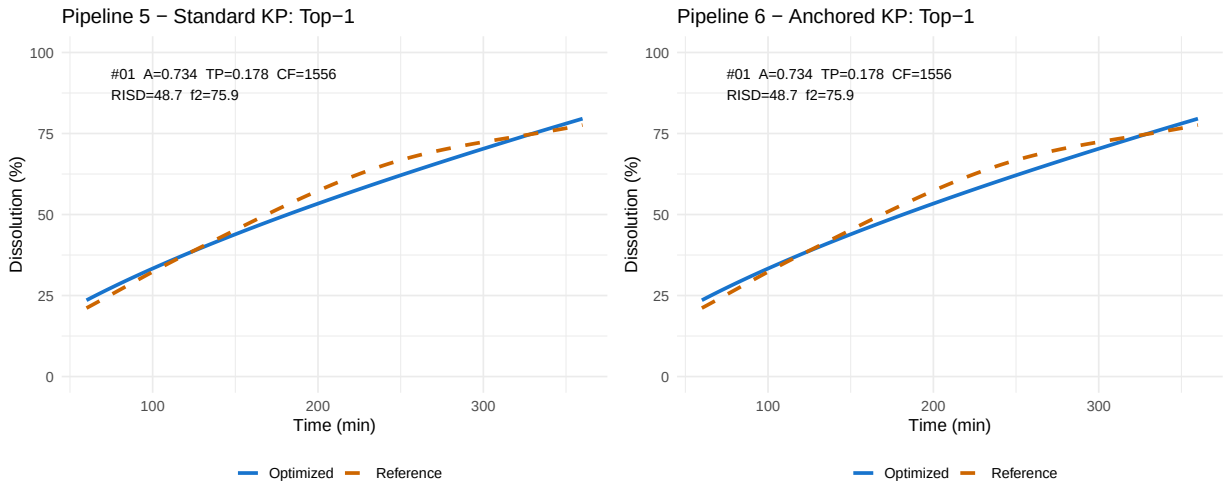


Figure 4.6: Dataset A: optimized dissolution curves under the KP pipeline.

Table 4.1: Best optimized candidate per pipeline for Dataset A under the full model, including standardized formulation variables

Pipeline	D_{RISD}	f_2	Polymer_A	Polymer_B	Total_Polymer	Compression_Force
Standard Weibull	18.70	89.4	0.7393	0.2607	0.1430	2416
Anchored Weibull	18.70	89.4	0.7393	0.2607	0.1430	2416
Anchored FPCA	21.16	86.8	0.7265	0.2735	0.1423	2429
Standard FPCA	21.52	88.2	0.7265	0.2735	0.1423	2429
Standard KP	48.68	75.9	0.7342	0.2658	0.1780	1556
Anchored KP	48.68	75.9	0.7342	0.2658	0.1780	1556

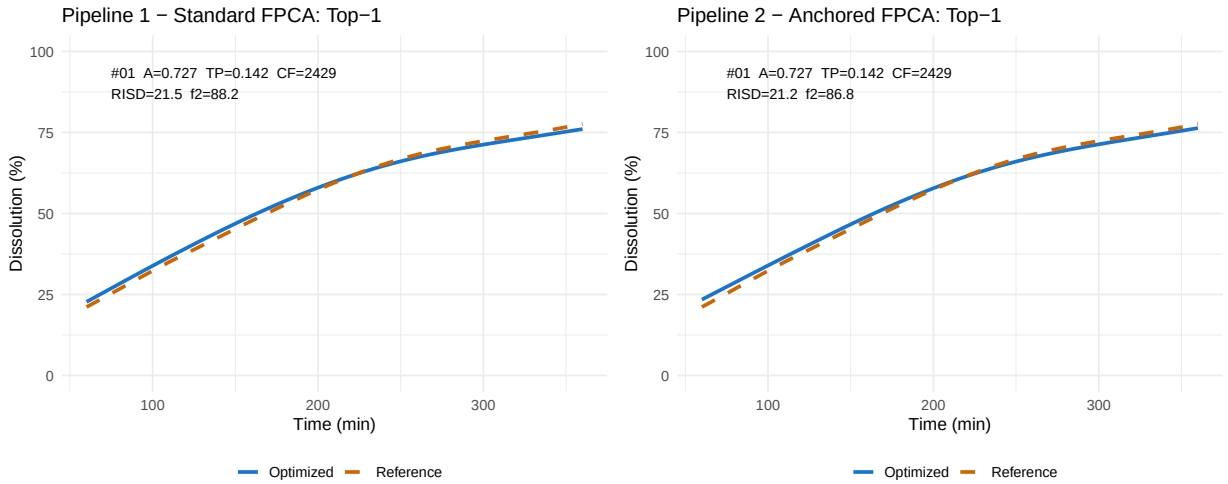


Figure 4.7: Dataset A: optimized dissolution curves under the FPCA (P-spline) pipeline.

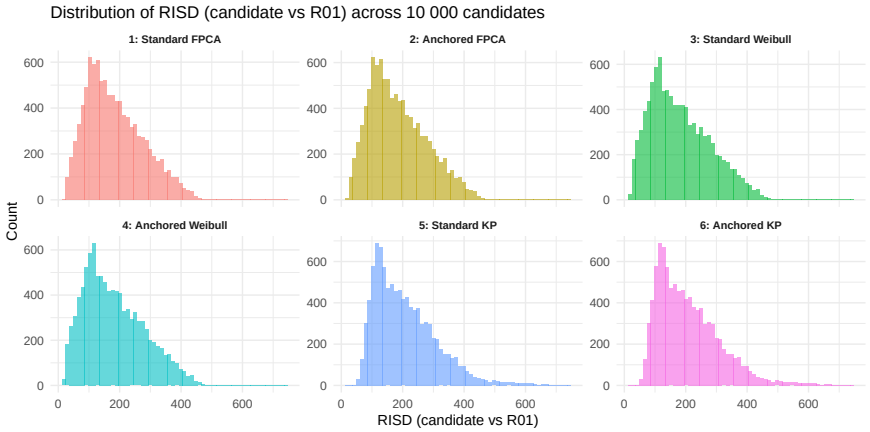


Figure 4.8: Distribution of D_{RISD} values across candidate formulations for each pipeline in Dataset A.

Table 4.2: Best optimized candidate per pipeline for Dataset A under variable selection, including standardized formulation variables.

Pipeline	VarSel	D_{RISD}	f_2	Polymer_A	Polymer_B	Total_Polymer	Compression_Force
Standard Weibull	forward	13.13	92.9	0.7623	0.2377	0.1241	1564
Anchored Weibull	backward	16.86	92.2	0.7267	0.2733	0.1354	2484
Anchored FPCA	backward	22.43	85.9	0.7265	0.2735	0.1423	2429
Standard FPCA	backward	22.62	87.3	0.7265	0.2735	0.1423	2429
Anchored KP	forward	39.31	75.0	0.7759	0.2241	0.1742	2398
Standard KP	forward	40.86	73.4	0.7610	0.2390	0.1752	2391

The original experimental study identified a preferred formulation based on laboratory testing criteria. The optimized formulations obtained under the curve-based framework differ from those reported in the original study, reflecting differences in objective criteria and optimization strategy.

4.3.4 Discussion for Dataset A

The empirical results for Dataset A suggest that the three-parameter Weibull model provides an appropriate structural representation of the observed extended-release profiles. The dissolution curves exhibit smooth, monotonic behavior with gradually decreasing release rates, a pattern well aligned with the parametric flexibility of the Weibull family. This structural compatibility likely explains the strong optimization performance observed for the Weibull-based pipelines.

When the full quadratic model is used, the anchored and standard parametric pipelines produce identical optimization results. This occurs because the predictor structure in Dataset A induces linear dependencies that effectively embed a constant component within the span of the design matrix. In particular, although no explicit intercept is included in the standard model, the formulation variables are compositional, with polymer fractions satisfying Polymer A + Polymer B $=$ 1. As a result, one of the composition variables implicitly acts as an intercept due to this linear constraint.

Under these conditions, the constant shift introduced by the anchored formulation (through reference differencing) can be absorbed by the regression coefficients of the predictors. Consequently, the standard and anchored parameterizations become algebraically equivalent when the full model is fit, leading to identical fitted values and identical optimized candidates.

After variable selection, however, some of these linear dependencies are removed as predictors are excluded from the model. Once the constant vector is no longer fully represented by the remaining predictors, the anchor shift can no longer be completely absorbed into the coefficient estimates. Consequently, the anchored formulation introduces a small but measurable difference in the fitted response surface, which explains the modest improvement in D_{RISD} observed for the anchored pipelines after variable selection. Further details are provided in Appendix B, Section B.2.1.

The FPCA pipeline yields competitive performance across evaluation metrics, demonstrating that a data-driven functional representation can effectively approximate the reference profile. In this dataset, FPCA achieves a level of curve matching comparable to the parametric approach, indicating that the dominant features of the dissolution behavior are well captured within a low-dimensional functional representation. The KP model performs comparatively worse, particularly in later release regions, reflecting limitations of its power-law structure in capturing extended-release curvature.

Reference anchoring provides modest improvements in D_{RISD} across pipelines. The magnitude of improvement is relatively small, indicating that when the baseline model form is already well aligned with the underlying dissolution mechanism, anchoring primarily fine-tunes rather than substantially alters optimization outcomes. Consistent effects on formulation recovery are observed in simulation studies (Chapter 5).

These findings are consistent with the simulation results in Section 5.6.1, where parametric pipelines demonstrated strong performance under correctly specified Weibull-generating mechanisms. Together, the simulation and empirical results suggest that model adequacy plays a central role in determining whether additional functional flexibility or reference-based adjustments meaningfully influence optimization performance.

From a practical standpoint, Dataset A illustrates that when dissolution behavior conforms to a parsimonious parametric structure, a well-specified model may provide both interpretability and effective formulation recommendation, while functional approaches such as FPCA offer a flexible alternative that achieves comparable matching performance.

A limitation of the real data application is the small number of observed time points per dissolution profile. With only four sampling times, the smoothing step may be less stable, and the resulting functional representation may not fully capture the underlying curve dynamics.

In such settings, FPCA effectively operates on a low-dimensional representation of the data, and its interpretation as a continuous functional decomposition should be made with caution. Nevertheless, the framework remains applicable and provides a coherent structure for modeling and optimization.

4.4 Results for Dataset B

Dataset B consists of extended-release dissolution profiles, with a single observed curve per formulation. Unlike Dataset A, no designated reference product is provided. To enable reference-anchored modeling and optimization, formulation F12 is randomly selected as an internal reference and treated as $y_{\mathrm{ref}}(t)$ throughout the analysis.

Dataset B originally comprised 91 formulations. Consistent with the preprocessing procedure reported in Sousa (2025) [59], six formulations exhibiting pronounced early burst release behavior were excluded. The remaining 85 formulations were divided into training ($n = 58$), validation ($n = 18$), and testing ($n = 9$) subsets.

4.4.1 Data Overview

Compared with Dataset A, Dataset B exhibits greater structural variability across formulations and does not benefit from replicate averaging. As a result, modeling flexibility plays a more prominent role in reconstruction performance.

4.4.2 Curve Fitting Comparison

Figure 4.10–4.12 present representative fitted or reconstructed dissolution profiles for the nine independent test formulations across each modeling family, with the reference curve overlaid for comparison. To improve readability, two illustrative profiles are shown per modeling approach.

The full set of reconstruction results is deferred to Appendix B, Section B.1.2.

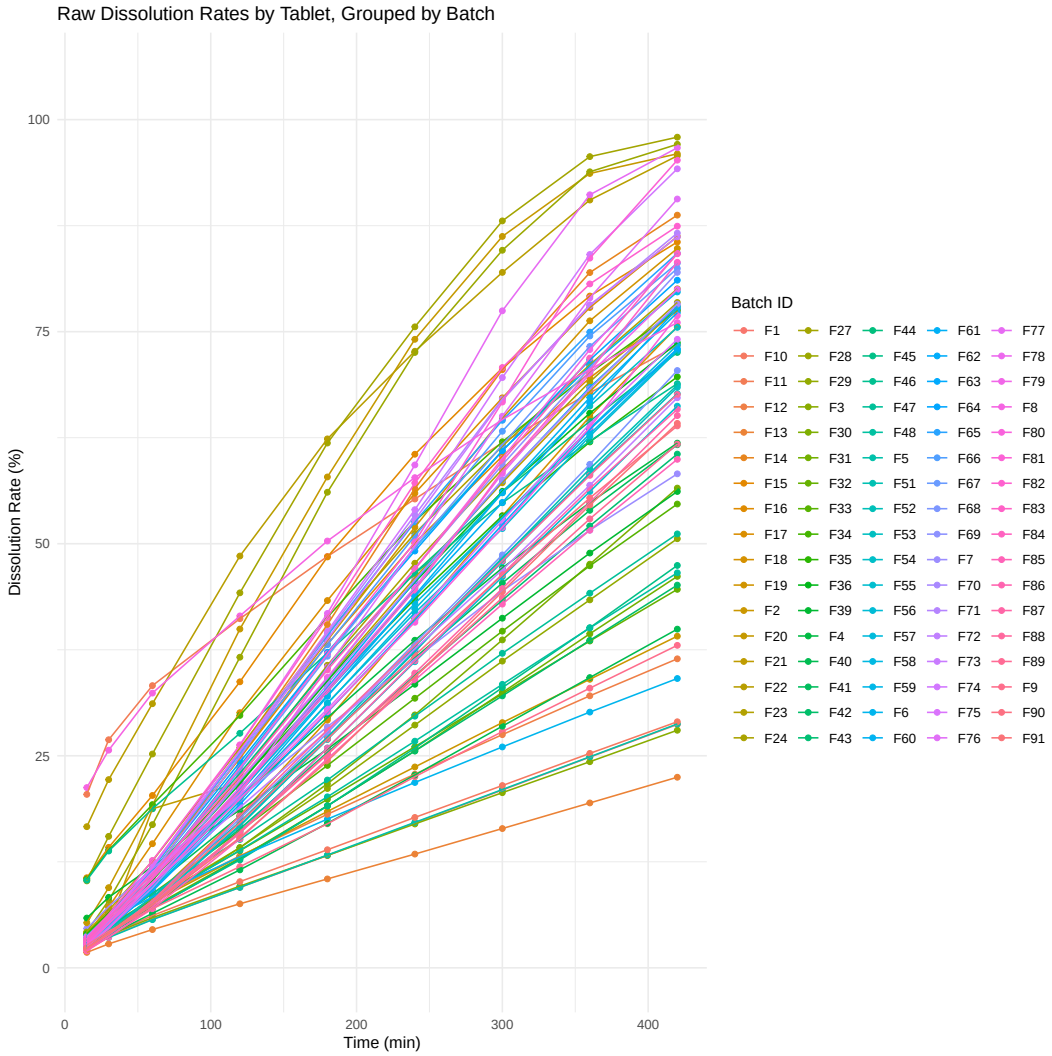


Figure 4.9: Raw dissolution profiles for Dataset B, where no designated reference formulation is available. For optimization, one observed profile is selected as a reference in practice, and each formulation is represented by a single measured curve, emphasizing reliance on modeling assumptions.

The reconstruction results reveal clearer separation between modeling families than observed in Dataset A. Because Dataset B contains only one observed curve per formulation and exhibits substantial heterogeneity, structural flexibility becomes critical.

The standard FPCA pipeline provides the strongest overall reconstruction performance on the test set. Its data-driven basis representation adapts naturally to diverse curvature patterns without imposing parametric restrictions, resulting in consistently lower D_{RISD} values and higher f_2 similarity for many formulations.

The reference-anchored FPCA model, in contrast, exhibits reduced reconstruction accuracy for formulations that differ substantially from the selected reference (F12). Since anchoring models deviations relative to the reference curve, predictions may partially reflect reference-specific structure, which can degrade reconstruction when true curve shapes are structurally distinct.

The Weibull model provides stable parametric fits and captures the overall monotonic release behavior effectively. However, its three-parameter structure limits flexibility in accommodating more complex curvature patterns present in certain formulations.

The KP model performs well when the underlying release trajectory resembles a power-law structure. For formulations deviating from this assumption, its restricted functional form leads to larger discrepancies.

Overall, Dataset B highlights the importance of structural flexibility when formulation-level heterogeneity is substantial. In this setting, the fully functional FPCA approach demonstrates a clear reconstruction advantage.

4.4.3 Optimization Outcomes

Optimization was conducted over the feasible formulation region $\mathcal{X}$ to identify

$$\hat{x} = \arg\min_{x \in \mathcal{X}} D_{\mathrm{RISD}}(x),$$

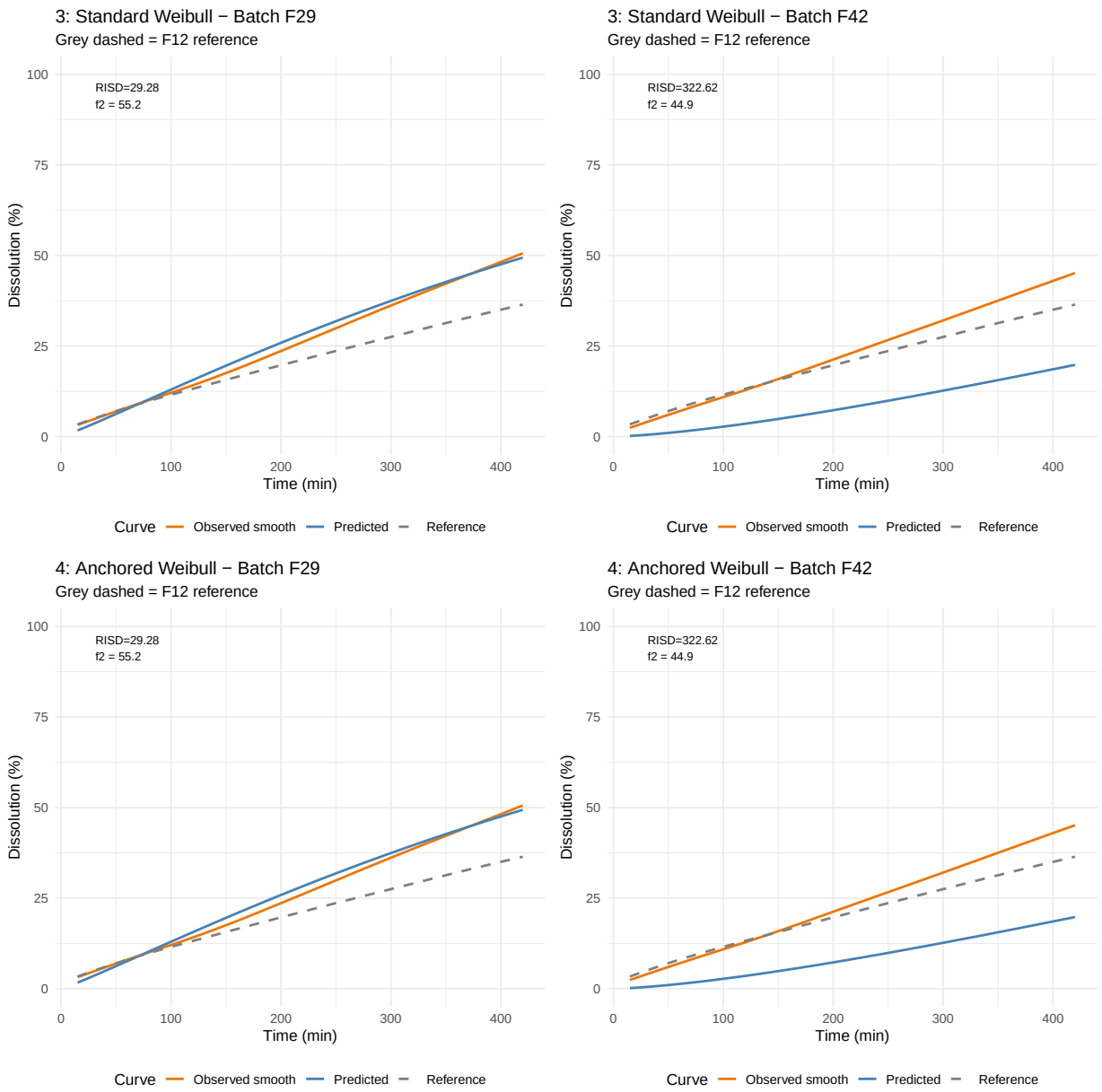


Figure 4.10: Dataset B: Fitted curves example under the Weibull model.

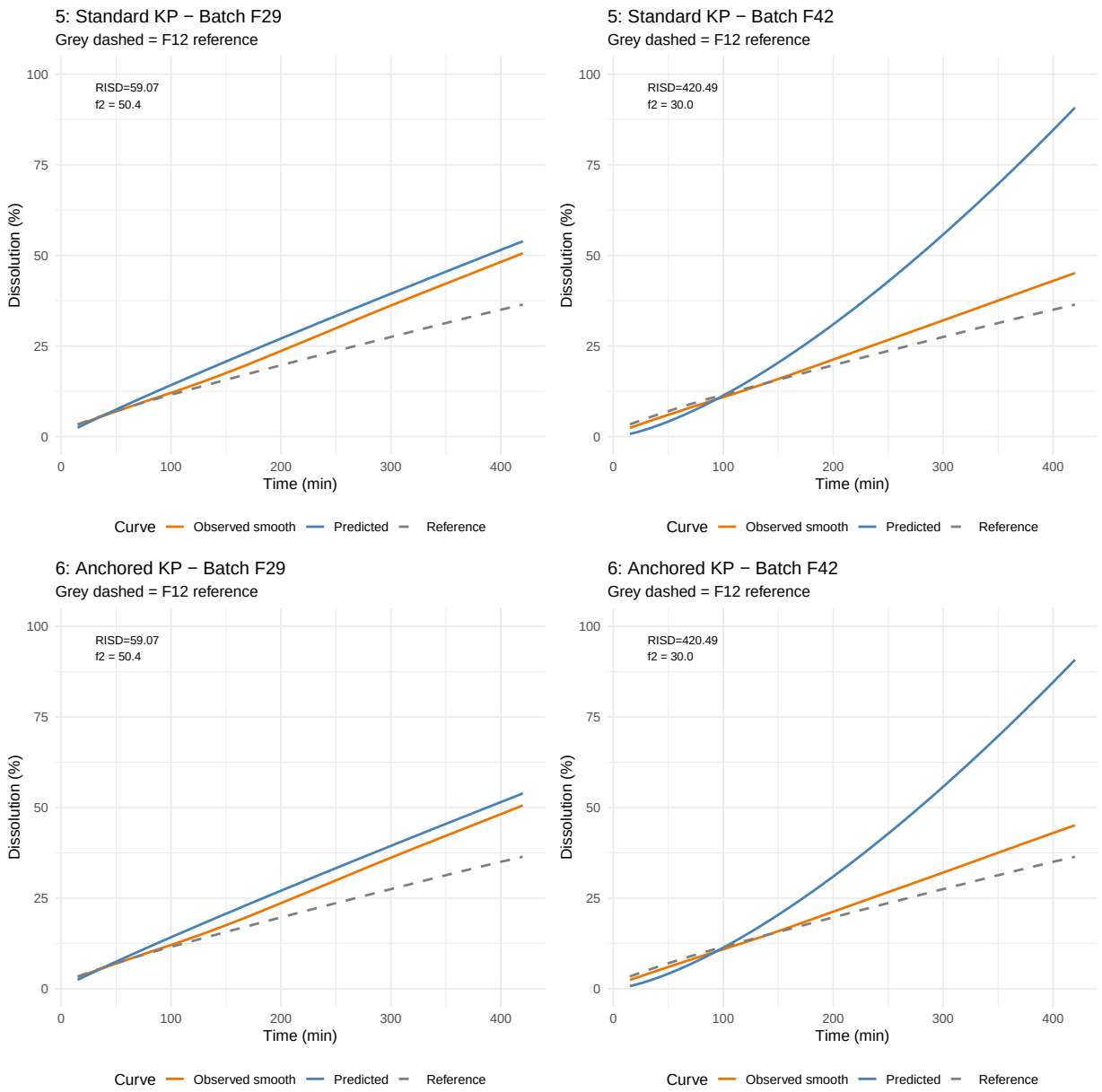


Figure 4.11: Dataset B: Fitted curves example under the KP model.

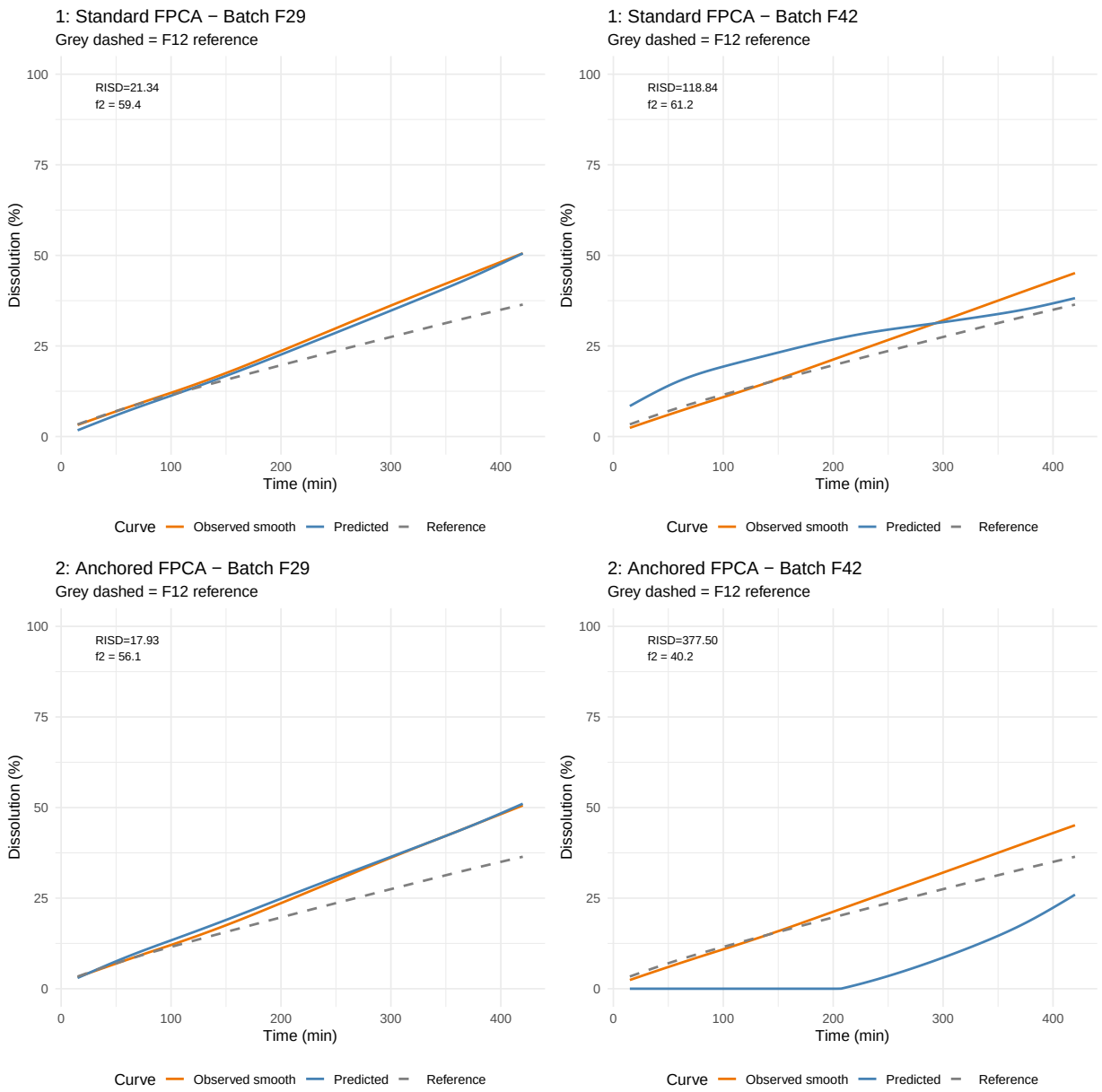


Figure 4.12: Dataset B: Fitted curves example under the FPCA (P-spline) model.

with f_2 reported as a secondary similarity measure. For Dataset B, formulation F12 was selected as the internal reference profile.

Figures 4.13–4.15presents the top-performing optimized candidates under each modeling pipeline, while Figure 4.16 summarizes the distribution of D_{RISD} values across the evaluated candidate set. Notably, the search-based framework identifies multiple near-optimal formulations with comparable performance, taking advantage of the non-uniqueness of the inverse mapping. These additional candidates are provided in Appendix B, Section B.2.2.

Numerical summaries of the optimal formulations and their corresponding D_{RISD} and f_2 values are provided in Table 4.3. Table 4.4 presents the same results with an additional column indicating the variable selection method used for each model.

Across pipelines, the KP-based models achieved the lowest D_{RISD} values under both the full model and variable selection settings. In the full model, the Standard and Anchored KP pipelines produced identical results ($D_{\mathrm{RISD}} = 8.89$, $f_2 = 96.6$), indicating the closest agreement with the selected reference formulation. Under variable selection, both KP variants further improved to $D_{\mathrm{RISD}} = 8.21$ with $f_2 = 97.4$, maintaining the strongest overall performance.

The FPCA and Weibull pipelines yielded moderately larger D_{RISD} values but still achieved high similarity scores, with all methods satisfying the conventional f_2 criterion ($f_2 > 90$). Under the full model, differences between standard and anchored variants were negligible for the KP and Weibull pipelines, reflecting the underlying model structure. When variable selection was applied, anchoring produced modest improvements in some cases, particularly for the FPCA pipeline.

Overall, these results indicate that the KP-based representation provides the best alignment with the selected reference profile in Dataset B, while FPCA and Weibull remain competitive alternatives. The consistency of performance across modeling choices, together with the relatively small differences between standard and anchored variants, suggests that model selection plays a more prominent role than anchoring in determining optimization outcomes for this dataset.

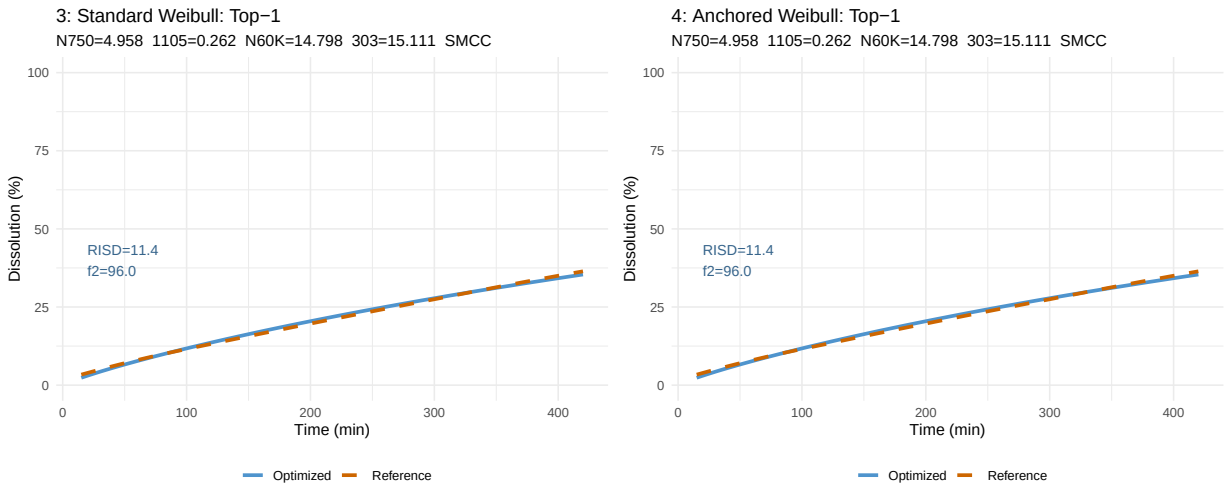


Figure 4.13: Dataset B: optimized dissolution curves under the Weibull pipeline.

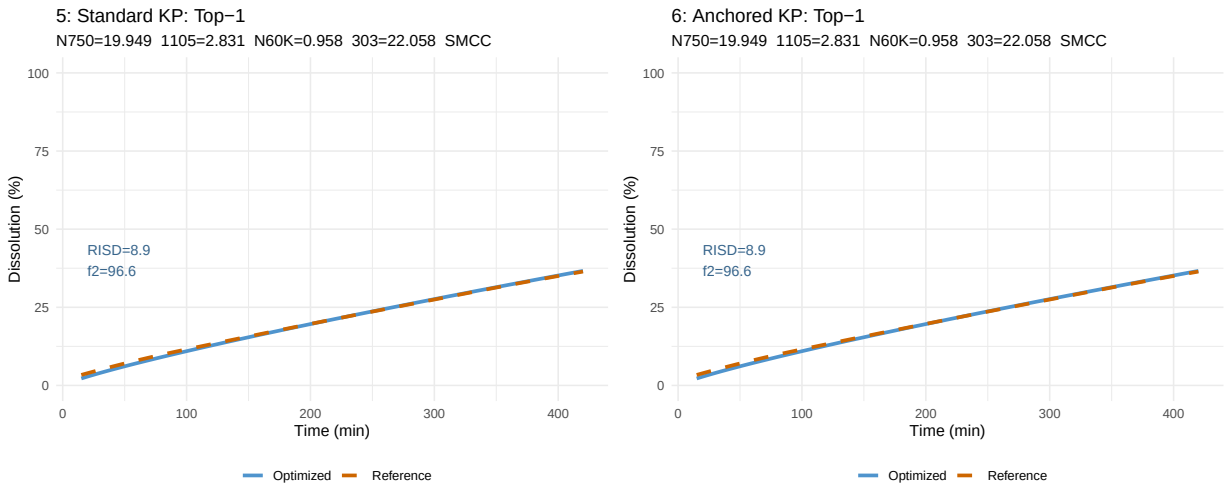


Figure 4.14: Dataset B: optimized dissolution curves under the KP pipeline.

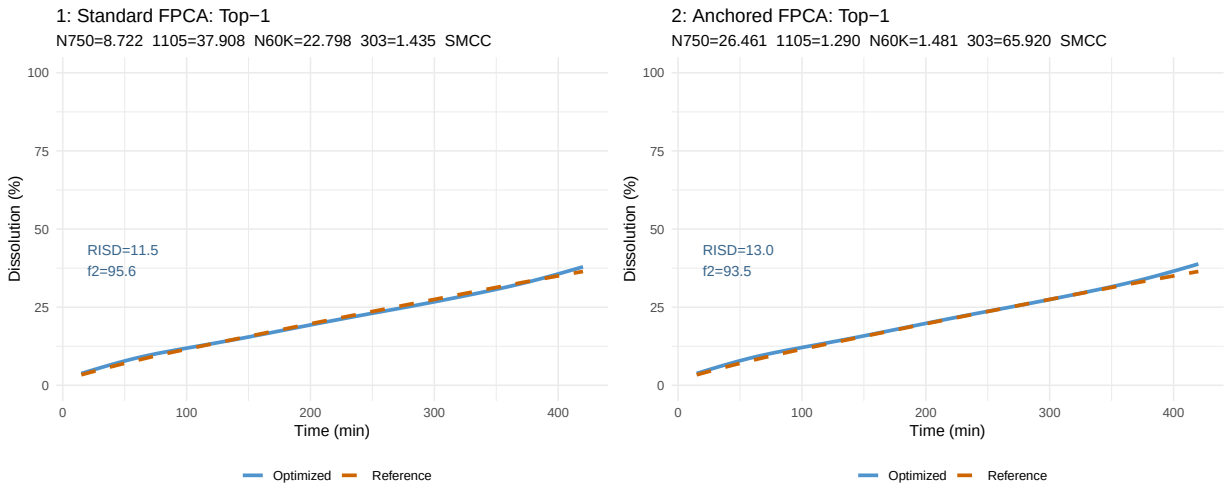


Figure 4.15: Dataset B: optimized dissolution curves under the FPCA (P-spline) pipeline..

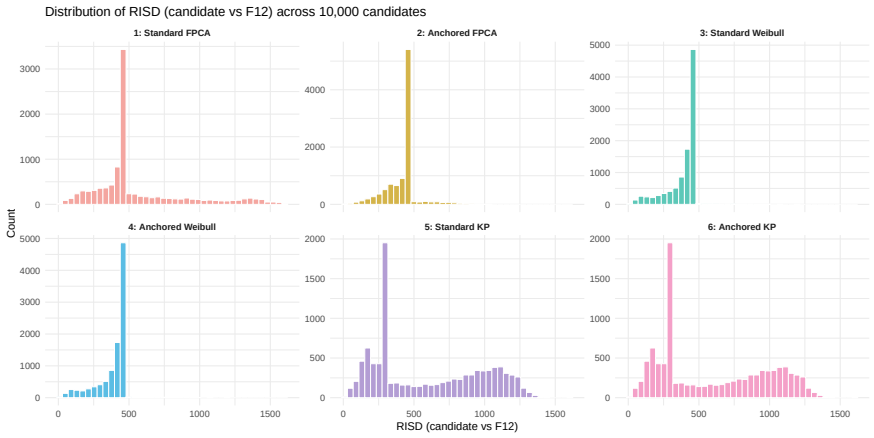


Figure 4.16: Distribution of D_{RISD} values across candidate formulations for each pipeline in Dataset B (reference = F12).

Table 4.3: Best optimized candidate per pipeline for Dataset B under the full model (reference = F12), ranked by D_{RISD}.

Pipeline	D_{RISD}	f_2	PEO_N750	PEO_1105	PEO_N60K	PEO_303	Diluent
Standard KP	8.89	96.6	19.949	2.831	0.958	22.058	SMCC
Anchored KP	8.89	96.6	19.949	2.831	0.958	22.058	SMCC
Standard Weibull	9.97	95.7	15.262	0.705	6.748	18.852	SMCC
Anchored Weibull	9.97	95.7	15.262	0.705	6.748	18.852	SMCC
Standard FPCA	11.32	94.3	31.108	29.608	14.572	2.204	SMCC
Anchored FPCA	12.95	93.5	26.461	1.290	1.481	65.920	SMCC

Table 4.4: Best optimized candidate per pipeline for Dataset B under variable selection (reference = F12), ranked by D_{RISD}.

Pipeline	VarSel	D_{RISD}	f_2	PEO_N750	PEO_1105	PEO_N60K	PEO_303	Diluent
Standard KP	forward	8.21	97.4	15.602	7.103	1.907	14.173	SMCC
Anchored KP	backward	8.21	97.4	15.602	7.103	1.907	14.173	SMCC
Standard FPCA	backward	11.49	95.0	38.538	10.134	20.411	4.408	SMCC
Anchored FPCA	backward	13.60	95.2	19.949	2.831	0.958	22.058	SMCC
Standard Weibull	backward	15.21	94.1	19.949	2.831	0.958	22.058	SMCC
Anchored Weibull	backward	15.21	94.1	19.949	2.831	0.958	22.058	SMCC

4.4.4 Discussion for Dataset B

In contrast to Dataset A, Dataset B exhibits greater structural heterogeneity across formulations, with more diverse curvature patterns and no experimentally designated reference product. Under this setting, model performance becomes more sensitive to alignment between the modeling family and the selected internal reference profile.

The flexibility of FPCA allows it to accommodate a wide range of curve shapes under such heterogeneous conditions. However, this flexibility may also lead to less stable behavior near the boundaries of the time domain, particularly in the tail region. This limitation can be addressed by incorporating shape constraints, such as monotonicity constraints or additional smoothness regularization within the FDA framework.

Under the full-model specification, the anchored and standard parametric pipelines produce identical optimization results for both the KP and Weibull models. This arises from linear dependencies in the predictor structure, particularly due to the inclusion of categorical diluent indicators.

These indicators introduce collinearity that effectively embeds a constant component within the span of the design matrix.

As a result, the constant shift introduced by the anchored formulation can be fully absorbed when the full set of predictors is used, rendering the anchored and standard parameterizations algebraically equivalent. Consequently, both approaches produce identical fitted values and identical optimized candidates under the full model.

After variable selection, this exact equivalence no longer necessarily holds because the standard and anchored models may select different predictor sets. In this case, both approaches still identify the same top-ranked optimized candidate, but differences appear among the remaining near-optimal candidates. Thus, the agreement under variable selection is limited to the best solution rather than the full optimization ranking. Further details are provided in Appendix B, Section B.2.2.

For the FPCA and Weibull pipelines, variable selection leads to slightly larger D_{RISD} values, indicating a modest loss in curve-matching accuracy. However, the anchored variants in these cases yield formulations that are closer to the true underlying composition, suggesting improved recovery of formulation structure despite the small increase in dissimilarity.

The strong performance of the KP-based pipelines suggests that several formulations in Dataset B exhibit release trajectories that are well approximated by power-law behavior relative to the selected reference formulation F12. When such structural compatibility exists, a parsimonious parametric model can provide effective optimization performance.

More broadly, the differing ranking patterns between Datasets A and B reinforce that no single modeling family uniformly dominates across empirical settings. Model adequacy depends on the structural characteristics of the dissolution profiles and on the choice of reference profile, particularly when no external ground-truth reference is available.

4.5 Chapter Summary

This chapter evaluated the proposed parametric and FDA-based pipelines using two empirical extended-release dissolution datasets with differing structural characteristics.

For Dataset A, where dissolution profiles exhibit smooth, monotonic behavior and a designated reference batch is available, the three-parameter Weibull model demonstrated strong structural alignment with the observed release mechanism. Under this setting, Weibull-based pipelines achieved the most favorable optimization performance. This result underscores that when the underlying functional form of the dissolution process is well understood and correctly specified, a parametric model can provide both interpretability and strong optimization accuracy.

However, Dataset B revealed a different pattern. With greater between-formulation heterogeneity and no externally defined reference product, model performance became more sensitive to structural assumptions and reference selection. In this setting, relative rankings varied across modeling families, and no single parametric structure uniformly dominated across formulations.

Taken together, these findings highlight a central distinction between modeling approaches. Parametric pipelines can perform exceptionally well when the correct functional form is known or closely approximated. In contrast, when the structural form of the dissolution curve is uncertain, misspecified, or highly heterogeneous across formulations, functional approaches such as FPCA provide consistently reliable reconstruction and competitive optimization performance without requiring prior assumptions about the governing equation.

Thus, while parametric models offer efficiency and interpretability under correct specification, FDA-based methods offer structural robustness and adaptability when the true release mechanism is unknown. The empirical results therefore reinforce the broader conclusion that model choice should be guided by the degree of structural knowledge available and the heterogeneity of the formulation space.

The next chapter presents a comprehensive simulation study designed to further evaluate the proposed pipelines under controlled data-generating mechanisms, allowing systematic assessment of robustness to model misspecification and validation of formulation recovery performance.

Chapter 5

Simulation Study

5.1 Objectives of the Simulation Study

The purpose of this simulation study is to evaluate the behavior and robustness of the proposed dissolution modeling pipelines under controlled conditions where the true formulation–curve relationship is known.

Four mechanistically distinct data-generating mechanisms are considered to represent a range of dissolution behaviors commonly encountered in practice:

- Three-parameter Weibull release,
- Logistic sigmoidal release,
- Korsmeyer–Peppas (power-law) release,
- Hixson–Crowell erosion-based release.

In addition, structural perturbations of the Weibull model, including mixture-based and mildly misspecified variants, are examined to evaluate sensitivity to parametric misspecification.

The mathematical forms used for each generating mechanism are specified in Section **??**.

The study is designed to assess:

- Robustness across different data-generating curve families,
- Sensitivity to parametric misspecification,
- The impact of reference anchoring on optimization accuracy,

- Formulation recovery performance quantified by squared deviation in formulation space (Formulation Recovery Error, FRE).

Because the formulation–curve relationship is fully specified in each simulation scenario, the framework enables direct evaluation of both curve-matching performance and formulation identifiability.

To ensure that the simulated dissolution profiles remain pharmaceutically realistic, the functional forms and parameter ranges used in the data-generating mechanisms were selected in consultation with pharmaceutical formulation expertise. The resulting simulated curves exhibit dissolution behaviors consistent with commonly observed release profiles, including monotonic release, plausible release rates, and saturation toward complete drug release.

5.2 Data-Generating Mechanisms

Four distinct dissolution-generating mechanisms are considered to represent a range of curve behaviors. In all cases, model parameters are determined by the formulation vector x through the quadratic mapping defined in Section 5.3.

5.2.1 Weibull-Generated Profiles

The three-parameter Weibull[25] mechanism is defined as

$$y_{\text{true}}(t; x) = a(x) \left(1 - \exp \left[- \left(\frac{t}{b(x)} \right)^{c(x)} \right] \right), \tag{5.1}$$

where $a(x)$ denotes the asymptotic release level, $b(x)$ controls the time scale, and $c(x)$ governs the curve shape. This formulation permits flexible concave and sigmoidal release patterns.

5.2.2 Logistic Sigmoidal Profiles

The logistic [46] generating mechanism is defined as

$$y_{\text{true}}(t;x) = 100 \cdot \frac{1}{1+\exp[-\left(\eta_0(x)+\eta_1(x)t\right)]}, \tag{5.2}$$

where $\eta_0(x)$ controls the location of the inflection point and $\eta_1(x)$ determines the steepness of the transition. This formulation corresponds to a standard logistic (sigmoidal) model used to describe growth processes and S-shaped response trajectories in nonlinear regression and dissolution modeling.

5.2.3 Korsmeyer–Peppas Profiles

The KP [22] mechanism is defined as a capped power-law model:

$$y_{\text{true}}(t;x) = \min\left(k(x)t^{n(x)},\, 100\right), \tag{5.3}$$

where $k(x)$ is a kinetic constant and $n(x)$ is the release exponent. The upper bound at 100 enforces the natural percentage release constraint. This mechanism captures near-linear and power-law release dynamics.

5.2.4 Hixson–Crowell Profiles

The Hixson–Crowell mechanism [18] is defined as

$$y_{\text{true}}(t;x) = 100\left[1-\left(1-k(x)t\right)^3\right], \tag{5.4}$$

where $k(x)$ controls the rate of surface-erosion–driven release. This mechanism typically generates gradual, approximately linear curve behavior over the observed time range.

5.3 Baseline Simulation Design

5.3.1 Formulation Space

Let $x \in \mathcal{X} \subset \mathbb{R}^p$ denote the formulation predictor vector. In the baseline setting, $p = 3$ predictors are considered. Training formulations are generated from a three-level factorial design, where each predictor takes coded levels $\{-1, 0, 1\}$. This produces a discrete set of design points used to fit each pipeline.

For the optimization stage, candidate formulations are searched over a continuous feasible region $\mathcal{X} = [-1, 1]^p$ using Latin hypercube sampling, allowing the recommended formulation $x^\star$ to take intermediate values between the original design levels.

5.3.2 True Parameter–Formulation Mapping

For each data-generating mechanism, model parameters are defined as deterministic functions of the formulation vector x. Specifically, each parameter $\theta_j(x)$ is generated on the log scale according to a full quadratic model:

$$\log\big(\theta_j(x)\big) = \beta_{0j} + \sum_{k=1}^{p} \beta_{kj} x_k + \sum_{k=1}^{p} \beta_{kkj} x_k^2 + \sum_{k<\ell} \beta_{k\ell j} x_k x_\ell, \tag{5.5}$$

where $\theta_j(x)$ denotes the j-th parameter of the data-generating curve (e.g., a, b, c for Weibull or k, n for KP).

The coefficient values $\{\beta\}$ are fixed across replicates and chosen to produce realistic dissolution behaviors while preserving smooth relationships between formulation variables and curve characteristics.

5.3.3 True Curve Generation

For a given formulation vector x, the true dissolution curve $y_{\text{true}}(t; x)$ is obtained by substituting the parameter values $\{\theta_j(x)\}$ into the corresponding data-generating function described in Section 5.2.

This construction establishes an explicit and known mapping from formulation space to curve space, enabling direct evaluation of formulation recovery performance.

In addition, the simulation settings are designed to span structurally distinct functional curve shapes and temporal dynamics, providing a controlled framework for assessing model performance under diverse functional behaviors and potential model misspecification.

5.3.4 Time Grid and Observation Model

Dissolution profiles are evaluated over discrete sampling grids

$$\mathcal{T}_{\text{obs}} = \{t_1, \ldots, t_m\},$$

representing typical laboratory observation schedules.

Two sampling designs are considered to evaluate model performance under different observation schemes:

- **Time grid A (IR-like sampling)**:

$$\mathcal{T}_{\text{A}} = \{5, 10, 15, 20, 30, 45, 60, 75, 90, 120\}$$

- **Time grid B (ER-like sampling)**:

$$\mathcal{T}_{\text{B}} = \{5, 30, 60, 90, 120, 180, 240, 360, 480, 600, 720, 900\}$$

Each data-generating mechanism is paired with one of the two sampling grids, and within each mechanism all competing modeling pipelines are estimated and optimized using the same observation grid $\mathcal{T}_{\text{obs}}$. While these grids are motivated by IR- and ER-like sampling schemes, the pairing with generating mechanisms is designed to prioritize diversity in curve shapes rather than enforce a strict classification. As a result, some profiles generated on the IR-like grid may

visually resemble ER behavior. This design does not affect the validity of the comparison, since all pipelines are evaluated under identical conditions within each mechanism, ensuring internal consistency for relative performance evaluation.

To compute the curve distance metric D_{RISD}, a dense evaluation grid

$$\mathcal{T}_{\mathrm{grid}}$$

of 500 equally spaced time points over the observation interval is used to approximate the integral numerically.

Observed dissolution values are generated at the tablet level to reflect within-batch variability.

For each formulation, six tablet-specific profiles are simulated.

For tablet r at time point t_j, the observation model is

$$y_{ir}(t_j; x) = y_{\mathrm{true}}(t_j; x) + \varepsilon_{ir}(t_j), \tag{5.6}$$

where the error process $\varepsilon_{ir}(t_j)$ follows an autoregressive model of order one (AR(1)) [66] across time:

$$\varepsilon_{ir}(t_j) = \phi\, \varepsilon_{ir}(t_{j-1}) + \eta_{ir}(t_j), \tag{5.7}$$

with $|\phi| < 1$ and innovations $\eta_{ir}(t_j)$ generated independently with mean zero and fixed variance. This structure induces realistic temporal correlation within individual tablet profiles.

Because dissolution is inherently monotone increasing, the simulated tablet-level curves are projected onto the space of monotone functions using the Pool-Adjacent-Violators Algorithm (PAVA) [2].

The batch-level observed profile is then obtained by averaging across tablets:

$$\bar{y}_i(t_j) = \frac{1}{6} \sum_{r=1}^{6} y_{ir}(t_j; x). \tag{5.8}$$

To the best of our knowledge, a simulation framework that integrates factorial formulation design, parametric mapping to dissolution models, and curve generation with correlated noise and monotonicity constraints has not been previously reported in the dissolution literature.

5.3.5 Reference Selection

For each simulation replicate, a reference formulation $x_{\text{ref}} \in \mathcal{X}$ is randomly sampled from the formulation space.

In the baseline scenarios, the reference profile is generated using the same observation model applied to all formulations: six tablet-level curves are simulated with AR(1) noise, monotonicity is enforced via PAVA, and the batch mean curve is computed. This averaged noisy curve serves as the observed reference profile used for optimization.

In the misspecification scenarios (Section 5.8), the reference curve is taken directly from the true data-generating function without added noise. This construction isolates structural model mismatch from stochastic observational variability.

5.3.6 Candidate Generation and Optimization

Because the inverse mapping from formulation space to curve space is not analytically invertible, optimization is conducted through finite candidate search.

For each simulation replicate, a set of candidate formulations

$$\left\{x^{(1)}, \ldots, x^{(N)}\right\} \subset \mathcal{X}$$

is generated using Latin hypercube sampling (LHS), which provides stratified coverage of the formulation space.

Each candidate formulation is evaluated under the fitted pipeline, and the curve distance $D_{\text{RISD}}(x)$ is computed relative to the reference profile. The formulation minimizing $D_{\text{RISD}}(x)$ is selected as the recommended solution, with $f_2(x)$ reported as a secondary metric.

The same candidate set is used across all pipelines within each replicate to ensure fair comparison.

5.3.7 Selection of the Number of FPCA Components

In the FPCA pipelines, the number of retained principal components was selected adaptively within each simulation replicate. For each replicate, FPCA was fit with an upper limit of $K_{\max} = 10$ components. Let $\{\lambda_k\}$ denote the estimated eigenvalues. The retained dimension K was chosen as the smallest integer such that the cumulative proportion of variance explained (PVE) exceeded a pre-specified threshold:

$$K = \min \left\{ k \leq K_{\max} : \frac{\sum_{\ell=1}^{k} \lambda_\ell}{\sum_{\ell=1}^{K_{\max}} \lambda_\ell} \geq 0.995 \right\}. \tag{5.9}$$

Because the complexity of the simulated curve families is moderate, the selected value of K is highly stable in the scenarios considered, and the retained dimension is typically $K = 2$ across replicates.

5.3.8 Pipelines Compared

The following six pipelines are evaluated:

- Weibull (Standard),
- Weibull (Reference-Anchored),
- KP (Standard),
- KP (Reference-Anchored),
- FPCA (Standard, P-spline smoothing),
- FPCA (Reference-Anchored).

All pipelines are supplied with the same observed data and the same candidate formulation set within each replicate, ensuring fair and controlled comparison.

The selected parametric pipelines represent commonly used and structurally distinct modeling assumptions in dissolution analysis. The three-parameter Weibull model provides flexible nonlinear and sigmoidal behavior, while the KP model captures power-law and near-linear release dynamics. Together, these models span a broad range of parametric assumptions typically encountered in practice.

The FPCA pipeline is included as a data-driven, model-agnostic alternative that does not impose a predefined functional form.

5.4 Evaluation Metrics

5.4.1 Curve Matching Accuracy

Curve similarity is evaluated using the primary criterion D_{RISD} and the secondary similarity factor f_2.

Throughout the simulation results, the label "RISD" in figures refers specifically to the curve-distance metric D_{RISD}, defined as the numerical integral of the absolute difference between the predicted and reference dissolution curves. This quantity represents the area *between* curves rather than the conventional area under a single dissolution profile.

5.4.2 Formulation Recovery Error

To assess the ability of each pipeline to recover the true underlying formulation, we define the *formulation recovery error* (FRE) as

$$\mathrm{FRE} = \|\hat{x} - x_{\mathrm{true}}\|^2, \tag{5.10}$$

where $\hat{x}$ denotes the recommended formulation and x_{true} is the true reference formulation.

Although this expression resembles a mean squared error, it represents squared deviation in formulation space rather than residual error from a fitted regression model. The FRE is averaged across simulation replicates to summarize recovery performance.

5.5 Illustration of Training Curves

To illustrate the variability induced by the three-level formulation design, Figures 5.1–5.4 display the 27 batch-level training curves generated under each data-generating mechanism. Each figure corresponds to one curve family (Weibull, Logistic, KP, and Hixson–Crowell, respectively).

These plots demonstrate the range of curve shapes produced across the formulation space prior to optimization. The systematic variation reflects the quadratic formulation–parameter mapping defined in Section 5.3.

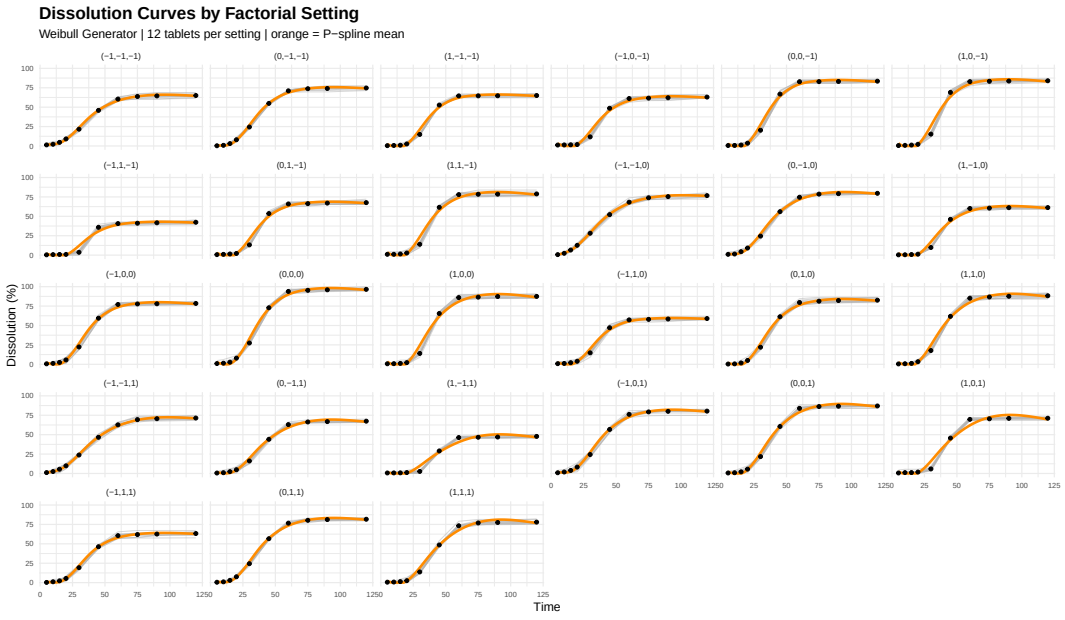


Figure 5.1: Training curves generated under the Weibull mechanism for the 3^3 formulation design.

The diversity of curvature patterns across generating mechanisms highlights the need for flexible modeling approaches that are not restricted to a single parametric form.

The logistic profiles exhibit a gradual, sigmoidal transition, representing a distinct functional shape that differs from standard parametric assumptions.

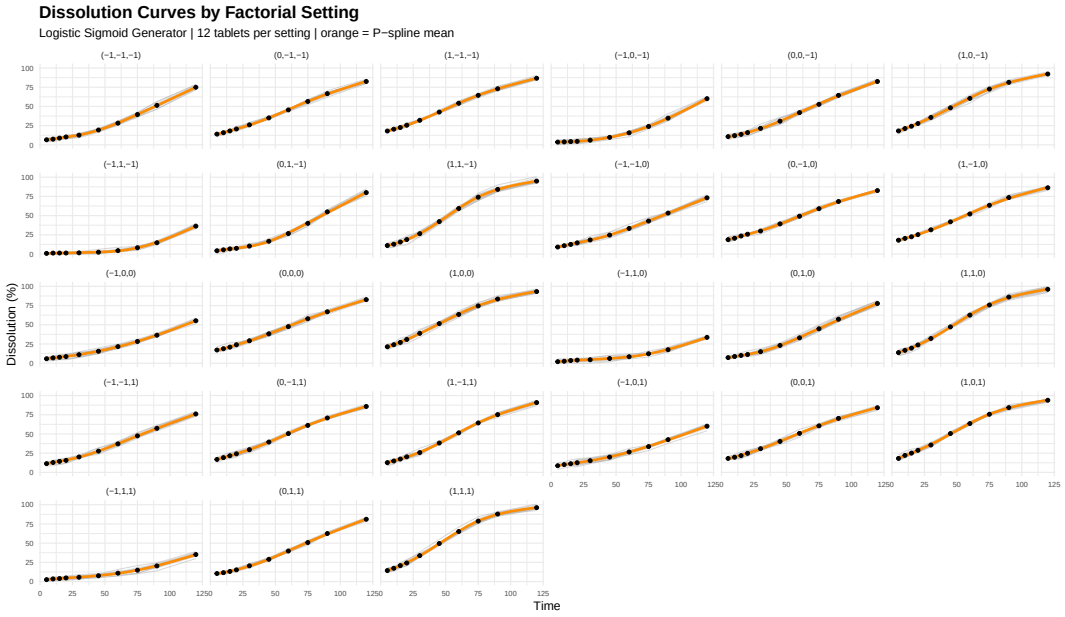


Figure 5.2: Training curves generated under the logistic mechanism for the 3^3 formulation design.

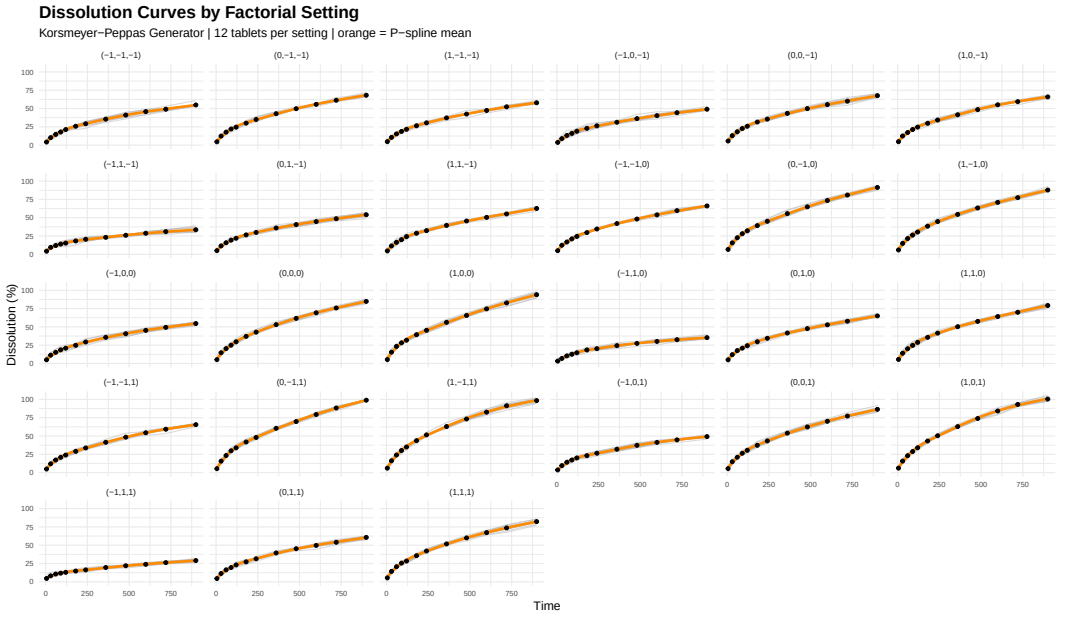


Figure 5.3: Training curves generated under the KP mechanism for the 3^3 formulation design.

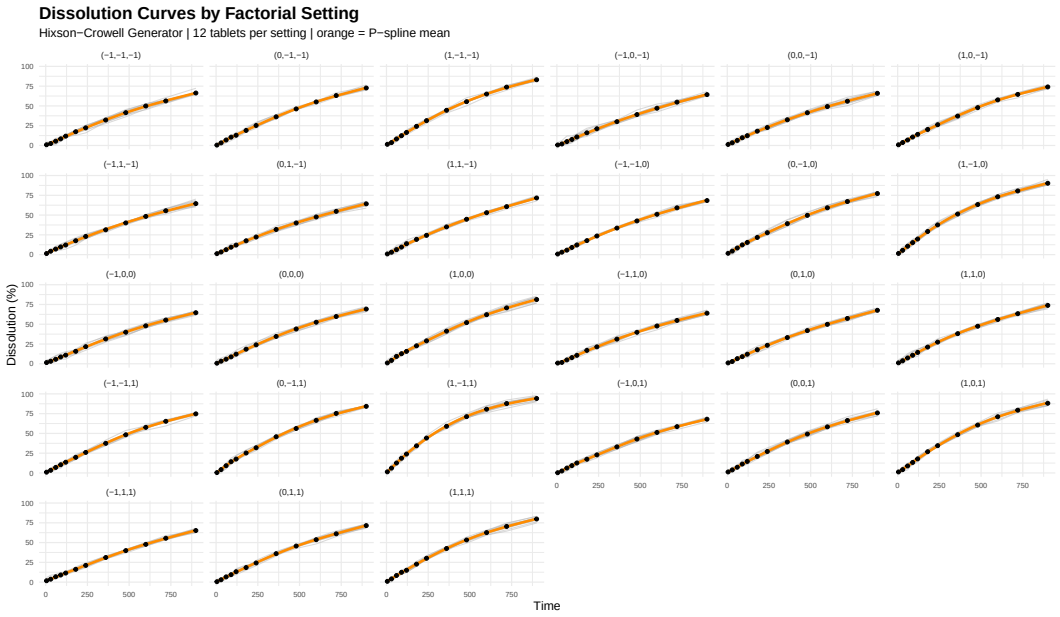


Figure 5.4: Training curves generated under the Hixson–Crowell mechanism for the 3^3 formulation design.

While the simulated curves were generated from prespecified parametric mechanisms and tuned to resemble realistic dissolution profiles, not every curve would necessarily be retained in an actual experimental development setting. In practice, profiles considered pharmaceutically implausible or operationally unsuitable may be excluded during early-stage screening.

In this simulation study, such curves are intentionally retained to provide a controlled evaluation of pipeline behavior under a broad range of structural conditions. This includes both well-specified settings, where the model aligns with the data-generating mechanism, and challenging misspecified scenarios. This design enables assessment of each method's robustness and its ability to adapt to diverse dissolution patterns.

5.6 Simulation Results: Robustness Across Curve Families

Performance is summarized using three primary outcomes: (i) curve distance D_{RISD}, (ii) similarity factor f_2, and (iii) formulation recovery error (FRE). Each plot compares all six pipelines under the same simulation setting.

5.6.1 Weibull-Generated Profiles

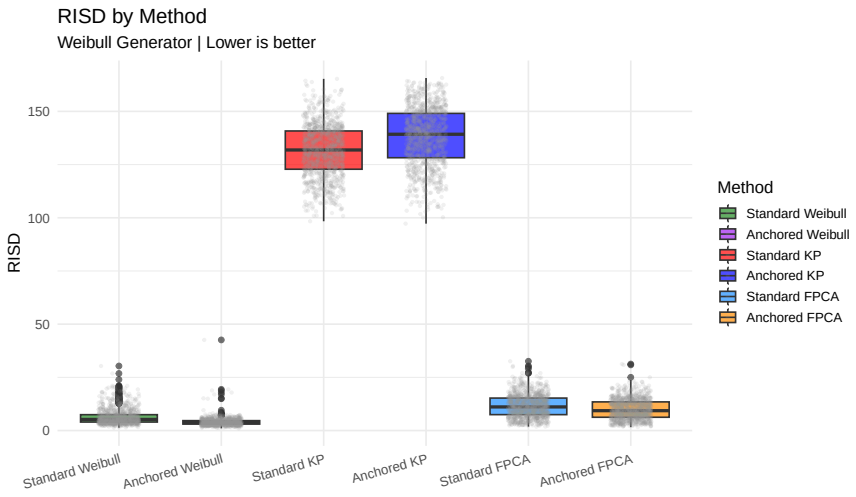
Curve Matching Performance. Figure 5.5a and Figure 5.5b summarize curve-level similarity using the area between curves (D_{RISD}) and the similarity factor f_2. Because the data-generating mechanism follows a Weibull model, the parametric Weibull pipeline provides the closest match to the reference dissolution profiles. Both the standard and anchored Weibull approaches produce very small D_{RISD} values and consistently high f_2 scores, typically well above the regulatory similarity threshold of 50.

In contrast, the KP-based pipelines perform poorly in this setting. Since the KP model assumes a power-law relationship that does not capture the sigmoidal Weibull structure, substantial model misspecification occurs. This results in very large D_{RISD} values and f_2 scores far below the similarity threshold.

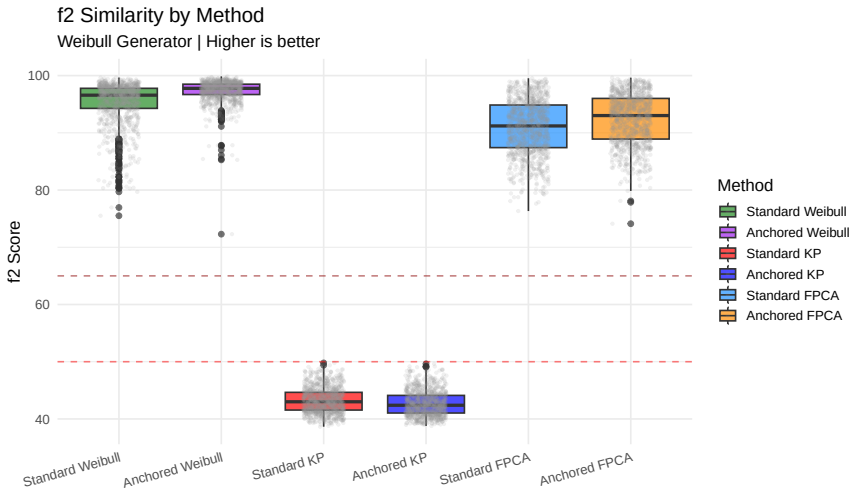
The FPCA pipelines provide an intermediate level of performance. Although FPCA does not assume a specific parametric form, its finite-dimensional representation introduces approximation error relative to the true Weibull model. Consequently, FPCA produces larger D_{RISD} values and slightly lower f_2 scores than the Weibull pipeline, but still maintains strong similarity in most simulations.

Formulation Recovery. Formulation recovery is evaluated using the formulation recovery error (FRE), defined as the squared distance between the estimated optimal formulation and the true generating formulation. The Weibull pipelines achieve the lowest FRE values, indicating that they most accurately recover the underlying formulation parameters when the parametric model is correctly specified.

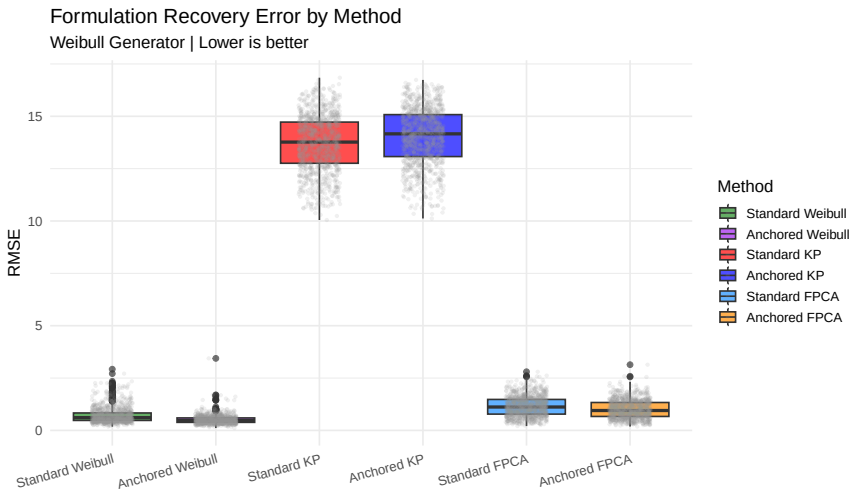
The KP pipelines exhibit extremely large FRE values, reflecting the inability of the KP model to represent the true Weibull-shaped dissolution curves. Even when curve-level similarity is partially achieved through parameter adjustment, the resulting formulation estimates are far from the true generating values.



(a) D_{RISD}



(b) f_2



(c) Formulation Recovery Error (FRE)

Figure 5.5: Simulation results for the Weibull data-generating mechanism.

The FPCA pipelines produce moderate FRE values, larger than those from the correctly specified Weibull model but substantially smaller than those from the KP pipelines. This behavior reflects the flexibility of FPCA, which can approximate a wide range of curve shapes, though the mapping from FPCA scores back to formulation space is less precise than in the correctly specified parametric model.

5.6.2 Logistic Sigmoidal Profiles

Curve Matching Performance. Figure 5.6 illustrates the dissolution profiles generated from the logistic sigmoid mechanism across the factorial design settings. The resulting curves exhibit a symmetric S-shaped trajectory with a relatively sharp transition region, which differs from the asymmetric structure implied by the Weibull model.

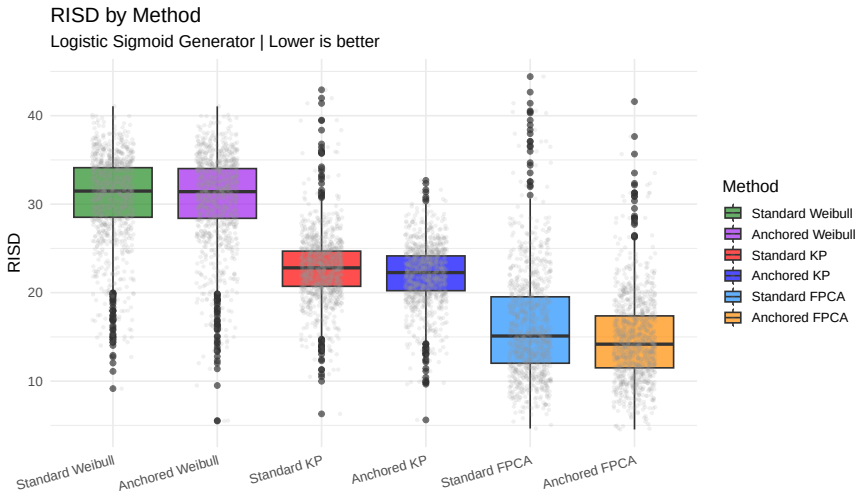
This structural mismatch is reflected in the similarity metrics shown in Figure 5.6a and Figure 5.6b. The Weibull pipelines produce substantially larger D_{RISD} values and lower f_2 scores compared with the other approaches, indicating that the Weibull model is unable to accurately capture the logistic curve shape under this design.

In contrast, the KP pipelines achieve improved curve similarity relative to Weibull. Although the KP model is not specifically designed for logistic curves, its flexible power-law structure provides a closer approximation to the logistic trajectory in this simulation setting.

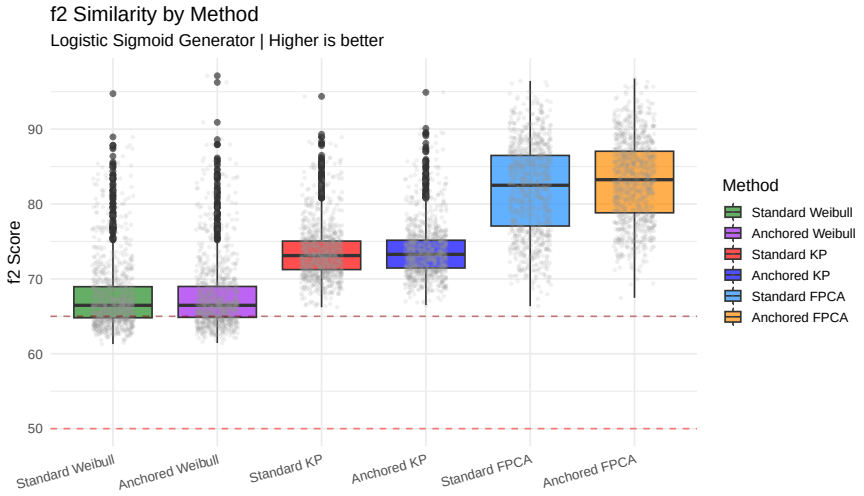
The FPCA pipelines yield the best overall curve matching performance. Because FPCA does not impose a predetermined functional form, it can flexibly represent the logistic sigmoid shape. Consequently, FPCA produces the smallest D_{RISD} values and the highest f_2 scores across most simulation replicates.

For the anchored KP pipeline, the log-normal bias correction $(\sigma^2/2)$ was omitted due to numerical instability and sensitivity of the KP model under the anchored specification.

Formulation Recovery. The formulation recovery error (FRE) results, shown in Figure 5.6c, follow a similar pattern. The FPCA pipelines achieve the lowest FRE values, indicating the most



(a) D_{RISD}



(b) f_2

(c) Formulation Recovery Error (FRE)

Figure 5.6: Simulation results for the Logistic data-generating mechanism.

accurate recovery of the underlying formulation parameters when the true dissolution mechanism is logistic.

The KP pipelines produce moderate FRE values, reflecting their ability to approximate the logistic curve shape to some extent but without fully capturing its structural characteristics.

The Weibull pipelines exhibit the largest FRE values. Because the Weibull model is misspecified for the logistic sigmoid curves, the optimization procedure compensates by adjusting formulation parameters in ways that reproduce the observed curve shape but deviate substantially from the true generating formulation.

5.6.3 Korsmeyer–Peppas Profiles

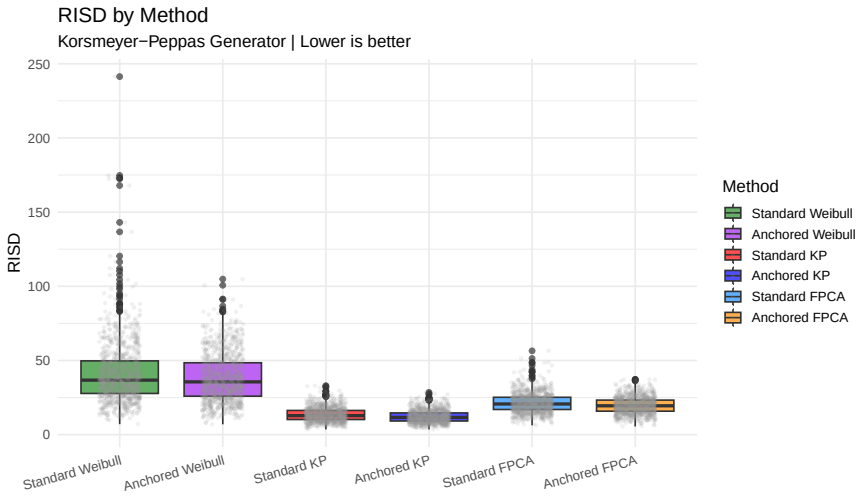
Curve Matching Performance. Figure 5.7 summarizes the simulation performance under the KP data-generating mechanism. The similarity metrics shown in Figure 5.7a and Figure 5.7b indicate that the KP pipelines achieve the strongest curve-matching performance in this setting.

Because the data-generating mechanism follows the KP power-law structure, the KP pipelines produce the smallest values of D_{RISD} and the highest f_2 scores across replicates. This result is expected under correct structural specification of the parametric model.

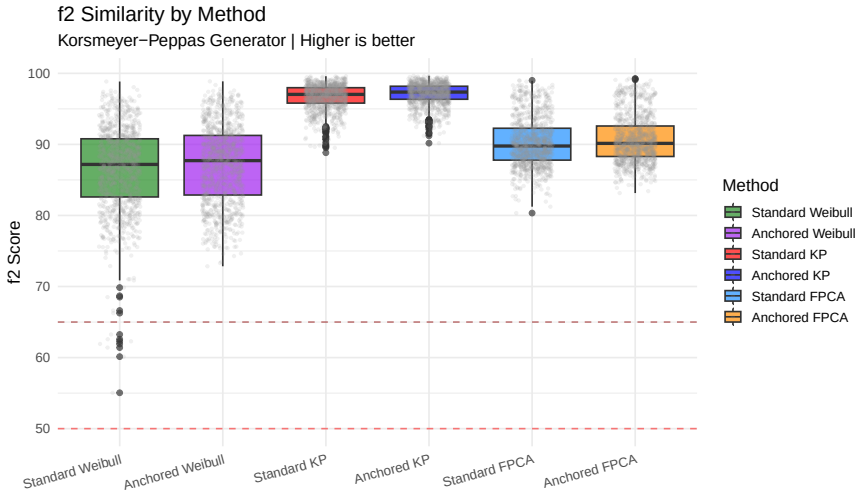
The FPCA pipelines demonstrate intermediate performance. Although FPCA does not assume a specific parametric form, its finite-dimensional functional representation introduces approximation error relative to the true KP mechanism.

In contrast, the Weibull pipelines exhibit larger D_{RISD} values and lower f_2 scores, reflecting structural mismatch between the Weibull functional form and the power-law behavior of the KP generator.

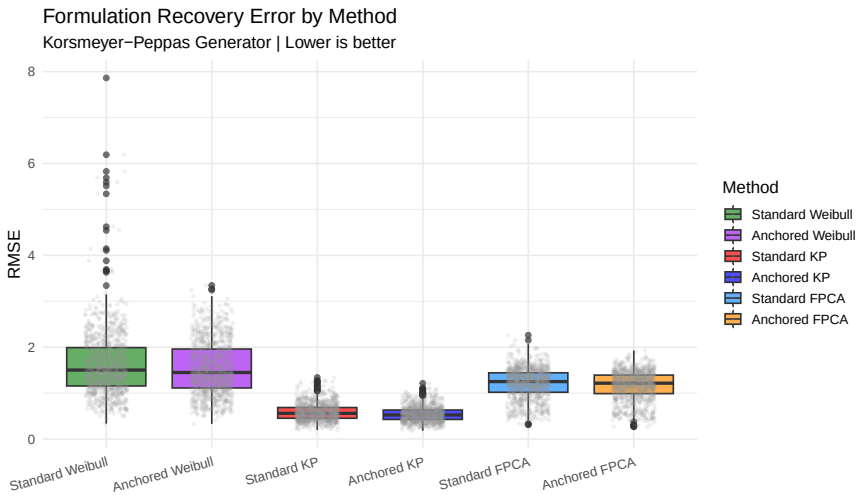
Formulation Recovery. The formulation recovery error (FRE) results, shown in Figure 5.7c, follow a similar pattern. The KP pipelines achieve the lowest FRE values, indicating accurate recovery of the underlying formulation–parameter relationship when the generating mechanism follows the KP structure.



(a) D_{RISD}



(b) f_2



(c) Formulation Recovery Error (FRE)

Figure 5.7: Simulation results for the Korsmeyer–Peppas data-generating mechanism.

The FPCA pipelines produce moderate FRE values, reflecting their ability to approximate the curve shape without explicitly modeling the power-law dynamics.

The Weibull pipelines exhibit substantially larger FRE values and greater dispersion across replicates, consistent with structural misspecification of the Weibull model under the KP data-generating mechanism.

5.6.4 Hixson–Crowell Profiles

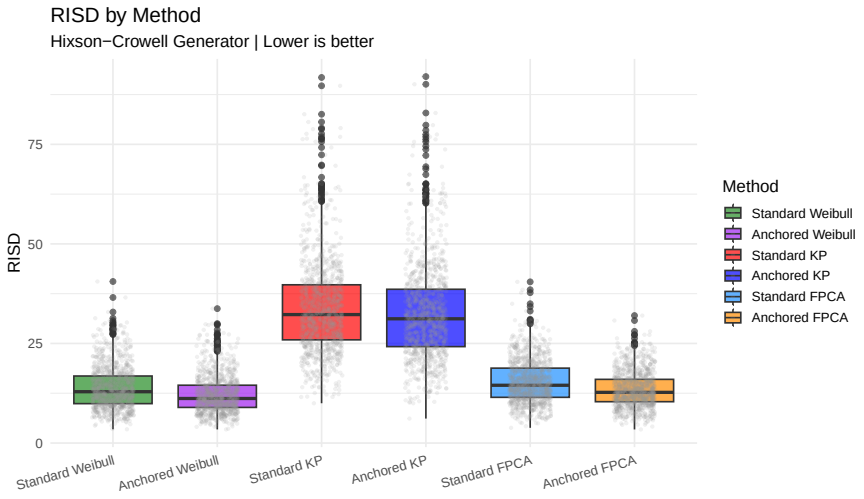
Curve Matching Performance. Figure 5.8 summarizes the simulation performance under the Hixson–Crowell data-generating mechanism. The similarity metrics shown in Figure 5.8a and Figure 5.8b indicate that the Weibull and FPCA pipelines achieve the strongest curve-matching performance in this setting.

Both methods produce relatively small values of D_{RISD} and consistently high f_2 scores across replicates. Although the Hixson–Crowell model is based on an erosion-driven release mechanism, its resulting curve shapes are sufficiently smooth and gradual that they can be approximated effectively by the flexible Weibull model and the data-driven FPCA representation.

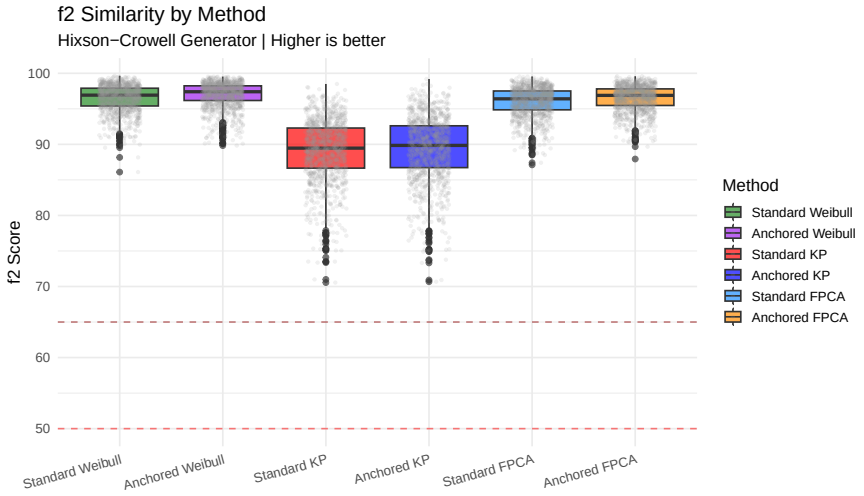
In contrast, the KP pipelines exhibit substantially larger D_{RISD} values and reduced f_2 scores. Although Hixson–Crowell curves may appear approximately linear over portions of the time domain, the underlying erosion-based mechanism does not follow a pure power-law relationship of the form t^n. Consequently, the KP model is unable to fully capture the curvature implied by the Hixson–Crowell formulation, resulting in poorer curve-matching performance.

Formulation Recovery. The formulation recovery error (FRE) results, shown in Figure 5.8c, follow a similar pattern. The Weibull pipeline achieves the lowest FRE values, indicating accurate recovery of the underlying formulation–curve relationship under the Hixson–Crowell generating mechanism.

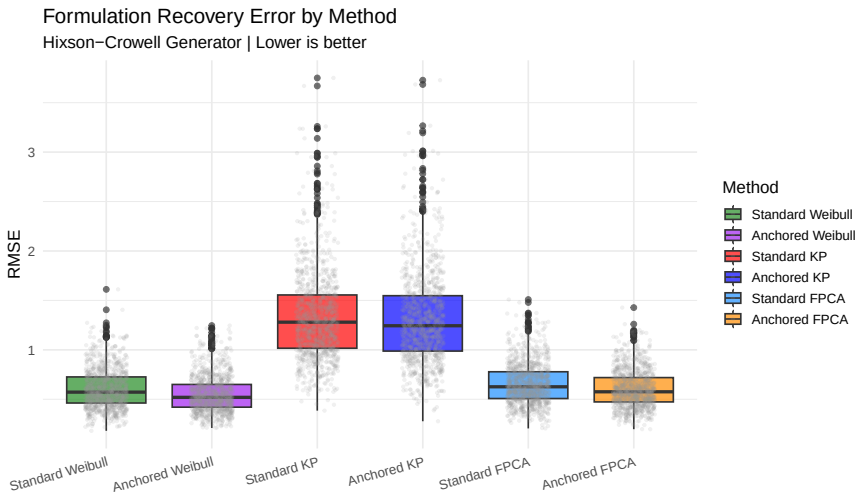
The FPCA pipelines exhibit slightly higher but relatively stable FRE across replicates, reflecting their ability to approximate the curve structure without imposing a specific parametric form.



(a) D_{RISD}



(b) f_2



(c) Formulation Recovery Error (FRE)

Figure 5.8: Simulation results for the Hixson–Crowell data-generating mechanism.

In contrast, the KP pipelines show substantially larger recovery error and greater dispersion, consistent with structural misspecification of the KP power-law model under the erosion-based release mechanism.

5.6.5 Overall Comparison Across Mechanisms

Table 5.1 summarizes the average performance of all six pipelines across the four data-generating mechanisms. Performance is evaluated using mean curve error, the standard deviation of error across replicates, the average curve distance D_{RISD}, and the similarity factor f_2.

Mechanism	Method	Mean Error	SD Error	Mean D_{RISD}	Mean f_2
Weibull	Standard Weibull	0.720	0.370	6.379	95.318
	Anchored Weibull	0.511	0.206	4.036	97.345
	Standard KP	13.687	1.358	131.423	43.234
	Anchored KP	14.019	1.356	137.932	42.712
	Standard FPCA	1.154	0.476	11.623	90.891
	Anchored FPCA	1.009	0.432	10.000	92.361
Logistic	Standard Weibull	4.380	0.847	30.687	67.797
	Anchored Weibull	4.374	0.856	30.539	67.840
	Standard KP	3.259	0.566	22.809	73.636
	Anchored KP	3.222	0.552	22.048	73.853
	Standard FPCA	2.148	0.750	16.261	81.918
	Anchored FPCA	1.994	0.641	14.731	83.075
Korsmeyer–Peppas	Standard Weibull	1.626	0.712	41.319	86.504
	Anchored Weibull	1.547	0.581	38.332	87.115
	Standard KP	0.588	0.184	13.530	96.662
	Anchored KP	0.541	0.163	12.199	97.106
	Standard FPCA	1.210	0.332	21.374	90.253
	Anchored FPCA	1.165	0.316	19.703	90.709
Hixson–Crowell	Standard Weibull	0.603	0.198	13.874	96.508
	Anchored Weibull	0.552	0.185	12.133	96.978
	Standard KP	1.337	0.461	33.912	89.081
	Anchored KP	1.313	0.472	32.566	89.330
	Standard FPCA	0.657	0.209	15.396	95.987
	Anchored FPCA	0.604	0.188	13.453	96.509

Table 5.1: Average simulation performance across all four data-generating mechanisms. Lower values indicate better performance for error and D_{RISD}, while larger values indicate better dissolution similarity (f_2).

5.6.6 Summary of Simulation Findings

The results in Table 5.1 reveal a consistent pattern in model performance. Parametric pipelines achieve the strongest accuracy when the assumed functional form aligns with the true data-generating mechanism. For instance, the Weibull pipelines perform best under Weibull-generated profiles, while the KP pipelines show superior performance when the underlying mechanism follows a KP power-law structure.

However, this advantage is conditional on correct model specification. When the parametric assumptions are violated, their performance deteriorates substantially. In contrast, the FPCA pipelines maintain stable and competitive performance across all scenarios. Although they may not always achieve the absolute best fit under ideal parametric conditions, they consistently avoid the large errors associated with model misspecification.

This contrast highlights an important implication: the gain from selecting a correctly specified parametric model is relatively modest, whereas the penalty for selecting an incorrect one can be substantial. Consequently, FDA-based approaches provide a more robust and reliable default strategy when the true dissolution mechanism is unknown.

5.7 Impact of Reference Anchoring

The impact of reference anchoring is evaluated across all simulation settings by comparing the standard and reference-anchored versions of each pipeline. Performance differences are assessed with respect to:

- Reduction in D_{RISD},
- Improvement in similarity factor f_2,
- Reduction in formulation recovery error (FRE).

Across the four generating mechanisms considered (Weibull, logistic, KP, and Hixson–Crowell), the qualitative effect of anchoring is remarkably consistent. In all scenarios, anchored pipelines

tend to produce slightly smaller D_{RISD} values and modestly improved formulation recovery relative to their non-anchored counterparts. The magnitude of improvement varies across mechanisms, but the direction of the effect is stable. This consistency suggests that anchoring acts primarily as a general stabilization device rather than a mechanism-specific correction.

Effect on Curve Matching

Across generating mechanisms, reference anchoring produces modest but systematic reductions in D_{RISD}. Although the magnitude of improvement is moderate, the direction of the effect is consistently favorable. Corresponding increases in f_2 are observed, indicating improved alignment with the selected reference profile.

The effect is somewhat more pronounced in settings involving structural misspecification, where the assumed parametric model differs from the true generating mechanism. In these cases, modeling deviations relative to the reference profile provides additional stabilization.

Effect on Formulation Recovery

In formulation space, anchoring yields slight but systematic reductions in FRE. While the improvements are not large, anchoring tends to reduce dispersion across replicates, suggesting improved stability of the optimization procedure.

Overall Assessment

Reference anchoring acts primarily as a stabilizing mechanism rather than a structural correction. It does not alter the relative ranking of pipelines under correct specification, nor does it fully compensate for strong model misspecification. However, it consistently improves reference alignment and modestly enhances formulation recovery, particularly under uncertain structural conditions.

5.8 Sensitivity to Parametric Misspecification

To evaluate robustness under structural model mismatch, an additional simulation scenario is considered in which the Weibull generating mechanism is replaced by a mixture-based construction. This mechanism preserves monotonic dissolution behavior while violating the assumption that a single three-parameter Weibull function governs release.

5.8.1 Mixture-Based Weibull Generator

In this scenario, the true dissolution curve is generated as a convex combination of two Weibull components [51]:

$$y_{\text{true}}(t;x) = 100\Big[(1-w)\Big(1-\exp\{-(t/b_1(x))^{c_1(x)}\}\Big) + w\Big(1-\exp\{-(t/b_2(x))^{c_2(x)}\}\Big)\Big], \quad (5.11)$$

where $w \in [0,1]$ controls the mixture weight. Both components share the same upper asymptote (100%), but differ in scale and shape parameters.

Although each component individually follows a Weibull form, their convex combination is generally not representable by a single three-parameter Weibull curve. This induces structural misspecification while preserving monotonic and physically plausible release behavior.

5.8.2 Results Under Mixture-Based Misspecification

This subsection summarizes the empirical behavior of the pipelines under the mixture-based Weibull generator. The mixture weight $w \in [0,1]$ controls the deviation from a single-Weibull structure: when $w = 0$ or $w = 1$, the generating curve reduces to a single Weibull component, while intermediate values of w produce a convex combination that is generally not representable by a single three-parameter Weibull curve.

Noise and monotonic enforcement. Within each formulation, tablet-level variability is introduced and the batch mean curve is computed. The mean curve is then projected onto the space of monotone functions using isotonic regression (PAVA) to enforce the physical constraint of non-decreasing dissolution. An illustration for representative mixture weights is shown in Figure 5.9.

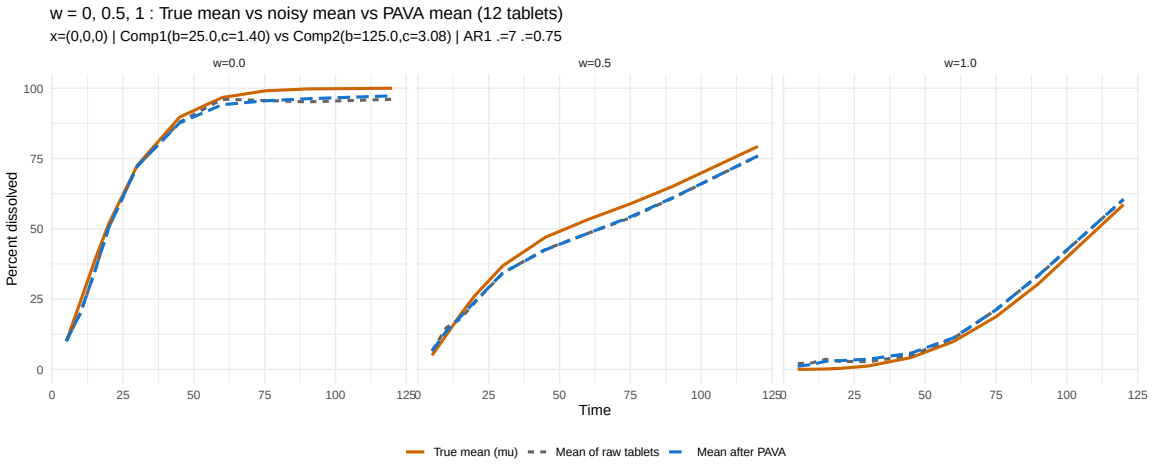


Figure 5.9: Illustration of the observation model under the mixture-based generator. For $w \in \{0, 0.5, 1\}$, the true noiseless mean curve (solid) is compared to the mean of simulated tablets (dashed) and the mean after PAVA monotone projection. The PAVA step restores monotonicity with minimal distortion to the underlying trend.

Shape evolution across mixture weight. To visualize the structural deviation induced by the mixture construction, Figure 5.10 overlays curves across a grid of mixture weights w for several representative formulation settings. As w increases, the curve transitions smoothly from the fast component toward the slow component while maintaining monotone release behavior.

Prediction accuracy for the diagnostic two-parameter Weibull model. To further assess whether the observed misspecification effect is driven by structural mismatch rather than by the particular choice of the three-parameter Weibull pipeline, the same prediction-accuracy analysis was first conducted using a single two-parameter Weibull fit as a diagnostic benchmark.

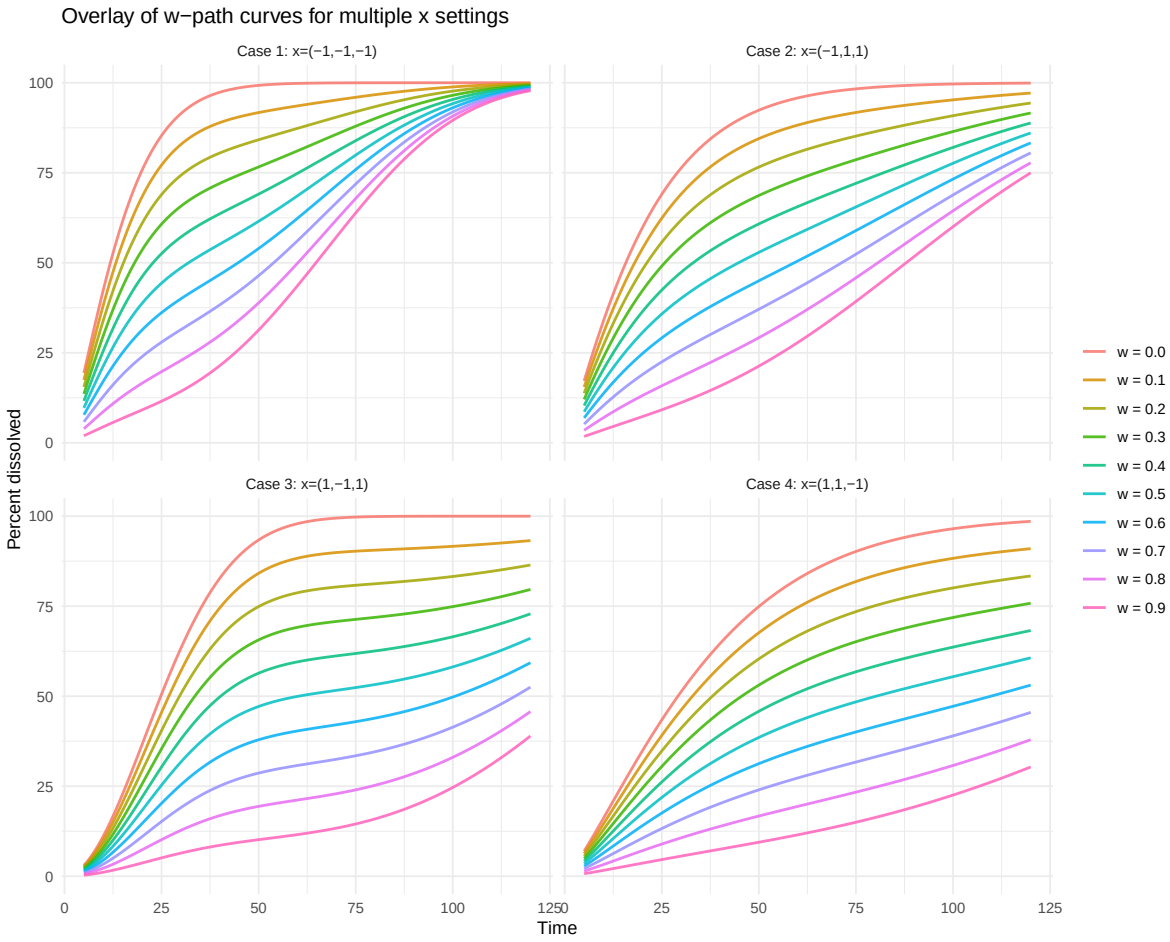


Figure 5.10: Overlay of mixture-based dissolution curves across mixture weights $w \in [0, 1]$ for several representative formulation settings x. The boundary cases ($w = 0$ and $w = 1$) reduce to single-component Weibull curves, while intermediate values of w generate curves that are not representable by a single three-parameter Weibull form.

Figure 5.11 summarizes prediction accuracy for the two-parameter Weibull model when evaluated against the noiseless reference curve across mixture weights w. Although the mixture components themselves follow two-parameter Weibull forms, their convex combination cannot generally be represented by a single Weibull curve. As a result, the diagnostic model exhibits clear performance degradation for intermediate mixture weights, confirming that the mismatch is structural rather than an artifact of the specific three-parameter pipeline.

A parallel comparison against the observed reference profile is shown in Figure 5.12. As expected, differences are attenuated when evaluation is performed against the noisy observed target rather than the noiseless truth. Nevertheless, the same qualitative pattern remains visible: the single

two-parameter Weibull model performs adequately near the boundary cases $w = 0$ and $w = 1$, but deteriorates for intermediate values of w.

Together, these two figures show that even when the component curves themselves are Weibull, a single Weibull model cannot generally recover their mixture over the full range of w.

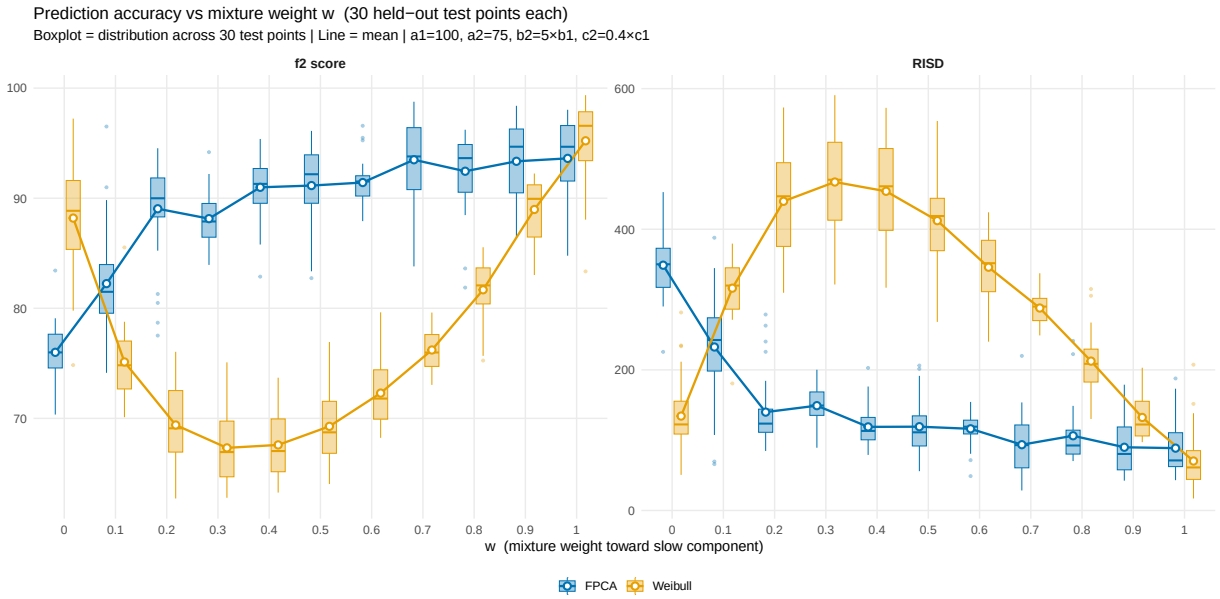


Figure 5.11: Diagnostic prediction accuracy for the two-parameter Weibull model as a function of mixture weight w under the mixture-based generator, evaluated against the noiseless reference curve. Accuracy is summarized using D_{RISD} and f_2.

Prediction accuracy for the three-parameter Weibull and FPCA pipelines. After establishing the structural limitation of a single Weibull model through the two-parameter diagnostic analysis, prediction accuracy was next compared for the standard three-parameter Weibull pipeline and the standard FPCA pipeline.

Figure 5.13 evaluates both methods against the noiseless reference curve across mixture weights w. Because the true curve reduces to a single Weibull component at the boundary cases $w = 0$ and $w = 1$, the Weibull pipeline performs strongly in those regimes. For intermediate mixture weights, however, the true curve is a convex combination of two distinct Weibull components and is generally not representable by a single Weibull curve. In this region, the Weibull pipeline

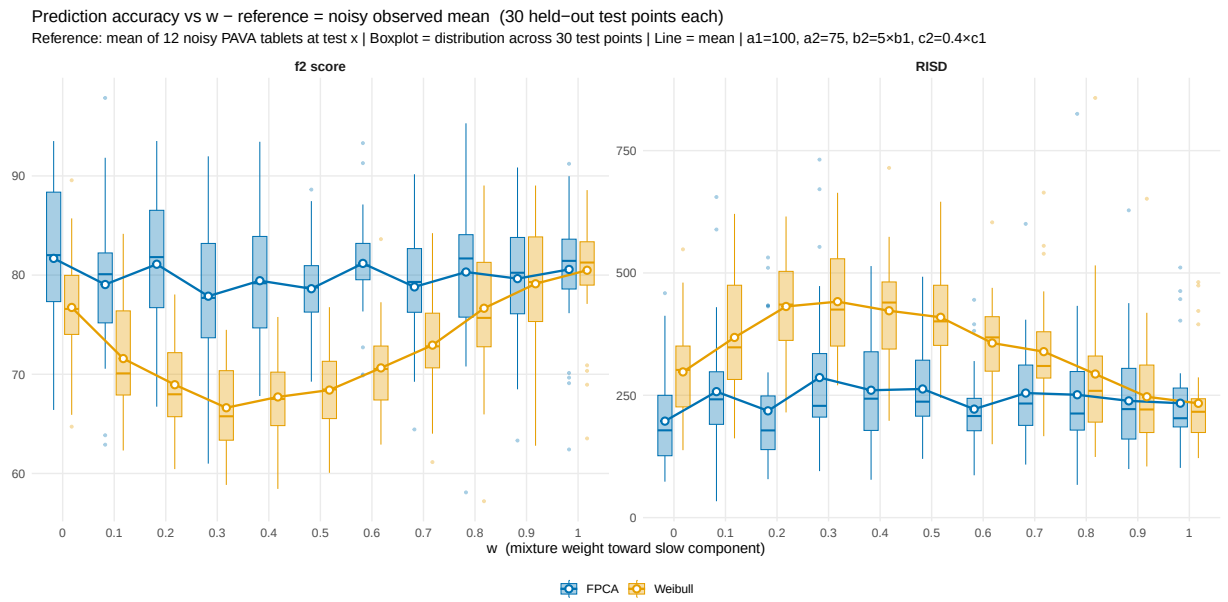


Figure 5.12: Diagnostic prediction accuracy for the two-parameter Weibull model as a function of mixture weight w under the mixture-based generator, evaluated against the observed reference curve after noise and monotonic projection. Accuracy is summarized using D_{RISD} and f_2.

exhibits moderate degradation, while FPCA remains comparatively stable due to its data-adaptive functional representation.

Figure 5.14 presents the corresponding comparison when evaluation is carried out against the observed reference curve obtained after tablet-level noise and monotonic projection. Because both methods are compared to the observed profile, absolute differences between pipelines are smaller than in the noiseless-truth comparison. However, the same qualitative pattern remains: the Weibull pipeline performs well when the generating mechanism remains close to a single Weibull form, but becomes more sensitive to mixture-induced shape distortion and observational perturbation, whereas FPCA retains more stable behavior across mixture weights.

Taken together, these two figures show that the three-parameter Weibull pipeline remains competitive near correctly specified boundary regimes, but FPCA provides greater robustness once the underlying release mechanism departs from a single parametric Weibull structure.

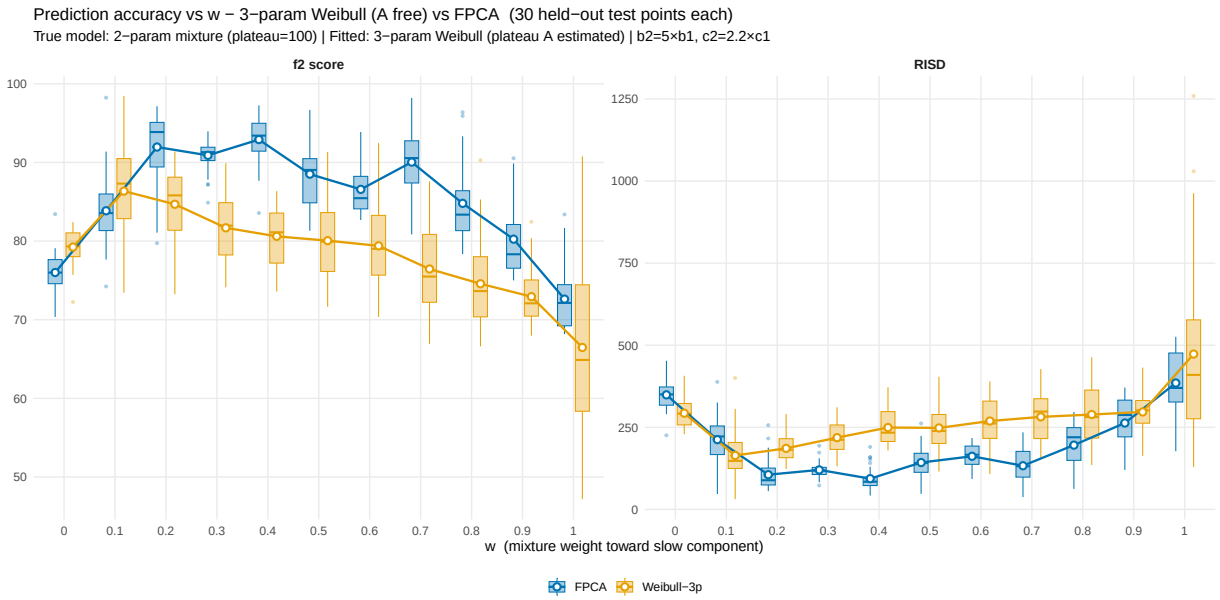


Figure 5.13: Prediction accuracy for the standard three-parameter Weibull pipeline and the standard FPCA pipeline as a function of mixture weight w under the mixture-based generator, evaluated against the noiseless reference curve. Performance is summarized using D_{RISD} and f_2 across held-out test settings.

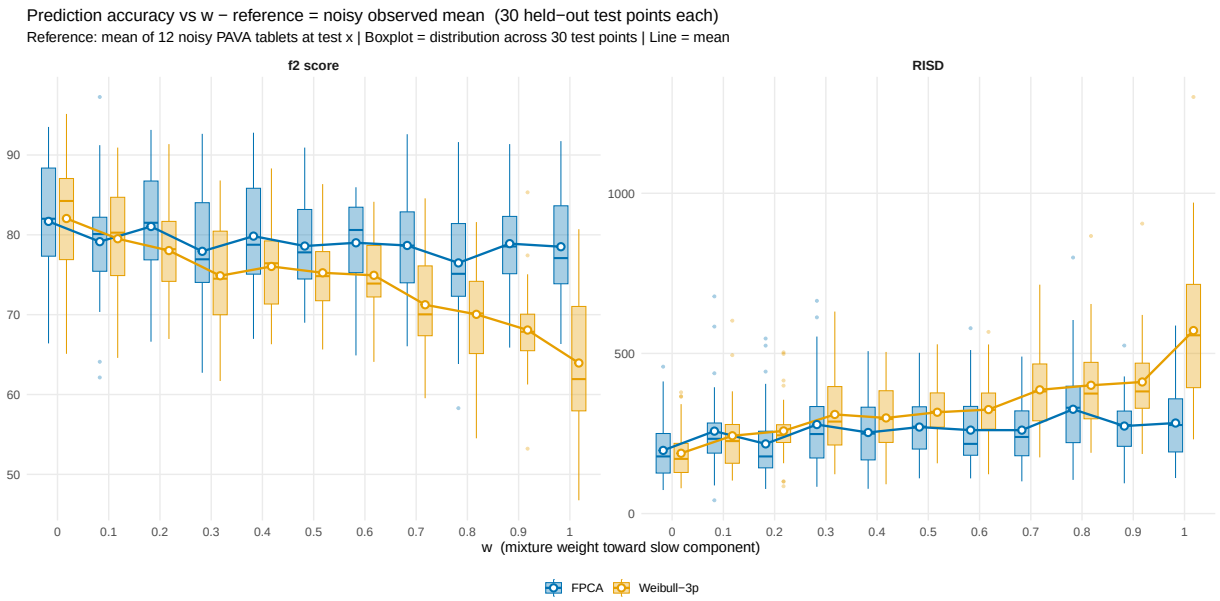


Figure 5.14: Prediction accuracy of the standard three-parameter Weibull and FPCA pipelines as a function of mixture weight w under the mixture-based generator, evaluated against the observed reference curve after noise and monotonic projection. Performance is summarized using D_{RISD} and f_2 across held-out test settings.

Summary. Overall, the mixture-based generator induces controlled structural misspecification. The parametric Weibull pipeline remains competitive near the boundary cases where the data-generating mechanism matches the assumed form, but it becomes moderately less accurate for intermediate mixture weights. FPCA maintains more stable curve-matching performance across mixture weights, consistent with the goal of a curve-agnostic framework under uncertainty in the underlying release mechanism.

5.9 Simulation Summary

This simulation study examined the comparative behavior of parametric and FDA-based dissolution modeling pipelines under controlled and fully specified data-generating mechanisms. By operating under known formulation–curve relationships, the framework enabled direct assessment of curve-matching accuracy and formulation recovery performance.

Across baseline generating mechanisms, parametric pipelines performed strongly under correct specification, while FDA-based approaches demonstrated competitive and stable behavior across diverse curve families. Under mixture-based misspecification, moderate degradation was observed for single-structure parametric models, whereas FPCA maintained comparatively robust performance.

Reference anchoring provided consistent, though moderate, reductions in D_{RISD} and slight stabilization of formulation recovery error, particularly in settings involving structural uncertainty.

Overall, the simulation results clarify the relative strengths and limitations of each pipeline under varying structural conditions. These findings provide a principled foundation for interpreting performance in real dissolution datasets, which are examined in the subsequent chapter.

Chapter 6

General Discussion

6.1 Overview and Synthesis of Findings

Across both simulated and empirical settings, all six modeling pipelines were successfully implemented and produced consistent and interpretable optimization results. These findings demonstrate that the proposed framework is effective in integrating parametric and functional approaches within a unified structure, and can be reliably applied across a range of dissolution scenarios.

The results show that model adequacy plays a central role in determining performance. Parametric pipelines achieve the best results when the assumed functional form closely matches the true dissolution mechanism (Section 5.8). However, this advantage depends critically on correct specification. When the assumed structure is incorrect, parametric performance can deteriorate substantially.

In contrast, FDA-based pipelines provide consistently stable and competitive performance across all settings. While they may not always achieve the absolute best fit under ideal parametric conditions, their performance remains reliable even when the underlying mechanism is unknown or misspecified.

This leads to an important trade-off. Selecting a correctly specified parametric model can yield improved accuracy, but requires prior knowledge of the true structural form. In practice, such knowledge is often unavailable, and incorrect specification can result in substantial performance loss. In comparison, the cost of using an FDA-based approach when a parametric model is correct is relatively small, while the benefit of its robustness under misspecification is significant.

From a decision-making perspective, these results suggest that FDA-based pipelines provide a reliable default strategy for dissolution modeling and formulation optimization, offering stable performance with minimal risk across diverse structural conditions.

Reference anchoring consistently reduces D_{RISD} in simulation settings and improves formulation recovery (Section 5.6.5), while its impact in empirical datasets is more moderate (Sections 4.3.3–4.4.3). Variable selection further enhances performance by reducing redundancy in the predictor space and improving model identifiability.

Overall, the proposed framework provides a flexible and effective approach for curve-based dissolution modeling and optimization. Within this framework, FDA-based components offer a stable foundation that ensures reliable performance, while parametric models can provide additional gains when their structural assumptions are well aligned with the underlying dissolution process.

6.2 Structural Model Assumptions and Model Adequacy

The results highlight the critical role of structural model assumptions in parametric modeling. When the parametric form is correctly specified, as in the Weibull-generating simulation(Section 5.6.1) and Dataset A(Section 4.3.3), the parametric pipelines achieve the lowest D_{RISD} values and the most accurate formulation recovery. This reflects the efficiency of parametric models when the underlying mechanism is well understood.

However, under structural misspecification or increased heterogeneity, parametric models become sensitive to model mismatch. This behavior is observed in mixture-based simulations(Section 5.8) and Dataset B(Section 4.4.3), where the Weibull functional form does not fully capture the underlying dissolution behavior.

In contrast, FDA-based approaches avoid reliance on a predefined functional equation and instead represent curves through data-driven basis functions. As a result, FPCA maintains stable and competitive performance across different generating mechanisms and empirical datasets.

This robustness has important implications. Although parametric models can be highly efficient under correct specification, their performance is sensitive to structural misspecification. In contrast, FDA-based approaches maintain stable accuracy across diverse settings. The resulting

trade-off is asymmetric: the performance gain from selecting a correct parametric model is relatively limited, whereas the performance loss from selecting an incorrect one can be substantial.

Consequently, FPCA-based approaches provide a more reliable and risk-robust choice for dissolution modeling, especially in settings where the true structural form is uncertain or varies across formulations.

6.3 Role and Impact of Reference Anchoring

Reference anchoring introduces a systematic adjustment to the modeling framework by centering predictions relative to a chosen reference profile. Across simulation scenarios, anchored pipelines consistently yield lower D_{RISD} values and improved formulation recovery, as measured by reduced FRE(Section 5.6.5). This suggests that anchoring stabilizes the optimization process by reducing systematic bias in the predicted response surface.

In empirical datasets, the effect of anchoring is more moderate. While improvements in D_{RISD} are generally observed, the magnitude of change depends on the alignment between the reference profile and the underlying structure of the formulation space.

Overall, anchoring serves as a refinement mechanism that enhances model performance without fundamentally altering the ranking of modeling approaches.

6.4 Comparison of Simulation and Real-World Behavior

The simulation study provides controlled settings in which the true data-generating mechanism is known, allowing direct assessment of model adequacy and formulation recovery. Under these conditions, parametric models perform best when correctly specified, while FPCA demonstrates strong robustness under misspecification and maintains consistently competitive performance even when parametric assumptions fail (Section 5.8).

These findings reinforce the role of FDA-based approaches as a default, as they provide reliable performance without requiring prior knowledge of the true dissolution mechanism.

The empirical results show similar patterns but with additional complexity due to unknown underlying mechanisms and the absence of ground-truth formulation parameters. Dataset A exhibits behavior consistent with Weibull-type release(Section 4.3.2), leading to strong parametric performance. In contrast, Dataset B exhibits greater structural heterogeneity(Section 4.4.2), increasing the likelihood of model misspecification when a single parametric form is assumed.

The agreement between simulation and empirical findings supports the validity of the proposed framework and demonstrates that the observed performance patterns are not dataset-specific but reflect general properties of the modeling approaches.

6.5 Curve-Based Optimization as a Design Tool

The proposed framework demonstrates that curve-based metrics such as D_{RISD} and f_2 can be effectively integrated into formulation optimization. By treating dissolution profiles as functional responses, the approach enables direct optimization of entire release trajectories rather than individual time points.

Variable selection further enhances this process by reducing model complexity and improving the stability of the optimization step. In the presence of collinearity(Section 4.3.3, 4.4.3), removing redundant predictors leads to more identifiable models and improved optimization performance.

These results suggest that curve-based optimization provides a flexible tool for formulation development, capable of accommodating both parametric and functional modeling strategies.

Prototype Interactive Visualization and Decision Support Tool

To demonstrate the applicability of the proposed framework, a prototype interactive visualization tool was developed. The tool allows users to explore the relationship between formulation variables and predicted dissolution profiles, and to compare candidate formulations against a selected reference curve.

By integrating model predictions with real-time visualization, the tool provides an intuitive interface for evaluating trade-offs between candidate formulations and supports decision-making in formulation development.

This implementation illustrates how the proposed pipeline can be extended beyond static analysis to serve as a design tool in pharmaceutical applications. A representative screenshot of the dashboard interface is provided in Appendix B, Section B.3.

6.6 Limitations

Several limitations should be acknowledged. First, the simulation settings, while diverse, may not capture all possible forms of dissolution behavior encountered in practice.

In addition, the choice of internal reference in Dataset B introduces potential variability in the optimization results. While anchoring improves consistency in many cases, its effectiveness depends on the representativeness of the selected reference profile.

Finally, the modeling framework focuses on curve-based similarity and does not explicitly incorporate additional manufacturing, stability, or process-related constraints, which may be relevant in practical formulation design.

Chapter 7

Conclusions and Future Directions

7.1 Conclusions

This dissertation developed and systematically evaluated a unified curve-based framework for dissolution profile modeling and formulation optimization. By integrating parametric (Weibull and KP) and functional (FDA-based) pipelines within a common optimization structure, the work provides a coherent basis for comparing modeling paradigms under both controlled simulation and empirical extended-release datasets.

The results demonstrate that model performance is fundamentally contingent on structural alignment. Parametric pipelines achieve superior accuracy and formulation recovery when the assumed functional form closely approximates the underlying dissolution mechanism. However, this advantage is conditional on correct specification. When structural assumptions are violated or the dissolution behavior is heterogeneous, parametric performance can deteriorate substantially.

In contrast, FDA-based approaches provide robust and consistently competitive reconstruction across diverse structural settings. While they may not always achieve the absolute best fit under ideal parametric conditions, their performance remains stable even when the underlying mechanism is unknown or misspecified.

This contrast reflects an important asymmetry in risk. The potential gain from selecting a correctly specified parametric model is relatively modest, whereas the penalty for selecting an incorrect one can be substantial. In comparison, the cost of adopting an FDA-based approach when parametric assumptions hold is small, while its robustness provides significant protection against misspecification.

From a practical perspective, these findings suggest that FDA-based pipelines offer a reliable default strategy for dissolution profile modeling and formulation optimization, particularly in settings where the true structural form is uncertain.

Reference anchoring provides a refinement mechanism that can stabilize optimization when a reliable and representative reference profile is available. However, anchoring does not compensate for structural misspecification and remains sensitive to reference quality.

Importantly, the proposed framework extends beyond curve matching to serve as a formulation design tool. By identifying multiple candidate formulations capable of achieving comparable dissolution behavior, the approach supports informed decision-making and expands exploration of formulation trade-offs within the feasible design space.

Collectively, these findings provide methodological guidance for selecting modeling strategies in dissolution profile modeling and formulation optimization more broadly. Although the empirical applications focused on extended-release datasets, the simulation study was designed to capture a wide range of curve behaviors. The observed performance patterns are consistent across these settings and reflect general properties of the modeling approaches rather than dataset-specific effects.

The choice between parametric and functional approaches should therefore be guided not only by predictive accuracy, but also by the level of structural uncertainty and the associated risk of model misspecification.

7.2 Methodological Contributions

This dissertation develops and evaluates a unified pipeline for dissolution profile modeling and formulation optimization. The main contributions are:

- The development of a unified analysis pipeline that integrates parametric and functional modeling approaches for dissolution profile comparison under both simulation and real-data settings.

- The introduction of a formulation recovery error (FRE) metric to quantify the ability of each method to recover the true underlying formulation in simulation studies.
- The implementation of a reference-anchored modeling strategy that improves stability and interpretability across different modeling approaches.
- A curve-based optimization procedure that enables direct identification of formulations matching a target dissolution profile.

7.3 Future Directions

Several avenues warrant further investigation.

First, optimization may be extended to incorporate preference-weighted objectives, allowing integration of predictor-space constraints such as manufacturing targets or robustness considerations.

Second, extending the framework to higher-dimensional formulation spaces is an important direction for future work. This includes studying settings with more formulation variables and complex interaction structures to assess computational efficiency and model stability under realistic industrial conditions.

Third, incorporation of alternative or multi-objective similarity metrics beyond D_{RISD} and f_2 may improve sensitivity to localized curve features or clinically meaningful release characteristics.

Fourth, adaptive or hybrid search strategies may improve computational efficiency while preserving transparency in candidate evaluation.

Fifth, incorporating formal goodness-of-fit assessment procedures may provide additional insight into how well the proposed pipelines capture observed dissolution behavior. While similarity metrics such as D_{RISD} and f_2 quantify agreement with a reference profile, they do not directly assess model adequacy. Future work could explore residual-based diagnostics, functional goodness-of-fit tests, or bootstrap-based procedures to evaluate the consistency between observed

and predicted curves. Such tools would complement the optimization framework by providing a more comprehensive assessment of model performance.

Sixth, the prototype dashboard developed in this work may be extended into a more general interactive platform for broader classes of dissolution and formulation data. Future development could incorporate support for multiple datasets, flexible model selection, user-defined reference profiles, and real-time visualization of candidate formulations under different optimization criteria. Such an extension would improve the practical accessibility of the proposed framework and strengthen its value as a decision-support tool in formulation development.

Finally, development of a dedicated R package for the proposed framework would facilitate reproducible implementation and broader adoption. Such a package could integrate modules for dissolution curve preprocessing, functional and parametric modeling, similarity evaluation, and candidate formulation optimization within a unified workflow.

Continued refinement along these directions has the potential to further strengthen the role of curve-based optimization as a methodologically grounded tool in pharmaceutical formulation science.

Bibliography

[1] A. Alvarado et al. In vitro biopharmaceutical equivalence of 5-mg glibenclamide tablets in simulated intestinal fluid without enzymes. *Dissolution Technologies*, 2021.

[2] Miriam Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and Edward Silverman. An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 26(4):641–647, 1955.

[3] B. Campisi, D. Chicco, D. Vojnovic, and R. Phan-Tan-Luu. Experimental design for a pharmaceutical formulation: optimisation and robustness. *Journal of Pharmaceutical and Biomedical Analysis*, 18(1-2):57–70, 1998.

[4] A. Choukri et al. Contribution of multivariate analysis to the in vitro dissolution profile for testing clopidogrel drugs similarity. *Dissolution Technologies*, 2022.

[5] T. J. Cleophas and A. H. Zwinderman. Functional data analysis (fda) advanced. In *Regression Analysis in Medical Research*, chapter 26. Springer, 2021.

[6] Paulo Costa and Jose Manuel Sousa Lobo. Modeling and comparison of dissolution profiles. *European Journal of Pharmaceutical Sciences*, 13(2):123–133, 2001.

[7] Dipankar Dey, Ritwik Ghosal, Kathleen R. Merikangas, and Vadim Zipunnikov. Functional principal component analysis for continuous non-gaussian, truncated, and discrete functional data. *Statistics in Medicine*, 2024.

[8] Aristides Dokoumetzidis, Vasiliki Papadopoulou, and Panos Macheras. Analysis of dissolution data using modified versions of noyes-whitney equation and the weibull function. *Pharmaceutical Research*, 23(2):256–261, 2006.

[9] Naihua Duan. Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association*, 78(383):605–610, September 1983.

[10] European Medicines Agency. Guideline on the investigation of bioequivalence (rev. 1), 2010. Guidance for Industry.

[11] R. W. Farebrother. Non-linear curve fitting and the true method of least squares. *The Statistician*, 47(1):137–147, 1998.

[12] Ronald A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.

[13] D. L. Galata, A. Farkas, Z. Könyves, L. A. Mészáros, E. Szabó, I. Csontos, A. Pálos, G. Marosi, Z. K. Nagy, and B. Nagy. Fast, spectroscopy-based prediction of in vitro dissolution profile of extended release tablets using artificial neural networks. *Pharmaceutics*, 11(8):400, 2019.

[14] K. Ghosal, S. Chakrabarty, and A. Nanda. Statistical modelling for controlled drug delivery systems and its applications in hpmc based hydrogels. In *AIP Conference Proceedings*, volume 1298, pages 219–224, 2010.

[15] A. T. C. Goh et al. Predicting drug dissolution profiles with an ensemble of boosted neural networks: A time series approach. *IEEE Transactions on Neural Networks*, 2003.

[16] M. C. Gohel et al. Assessment of similarity factor using different weighting approaches. *Dissolution Technologies*, 2005.

[17] Eitan Greenshtein and Ya'acov Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988, 2004.

[18] A. W. Hixson and J. H. Crowell. Dependence of reaction velocity upon surface and agitation. *Industrial & Engineering Chemistry*, 23(8):923–931, August 1931.

[19] Thomas Hoffelder. Equivalence analyses of dissolution profiles with the mahalanobis distance. *Biometrical Journal*, 2019.

[20] Ron S. Kenett and Chris Gotwalt. Functional data analysis and nonlinear regression models: An information quality perspective. *Quality Engineering*, 35(3):480–492, 2023.

[21] S. Kollipara et al. Simplified model-dependent and model-independent approaches for dissolution profile comparison for oral products: Regulatory perspective for generic product development. *AAPS PharmSciTech*, 2022.

[22] Richard W. Korsmeyer, Robert Gurny, Eric Doelker, Pierre Buri, and Nikolaos A. Peppas. Mechanisms of solute release from porous hydrophilic polymers. *International Journal of Pharmaceutics*, 15(1):25–35, May 1983.

[23] Pritam Kundu and Hans-Georg Müller. Intrinsic modeling of shape-constrained functional data, with applications to growth curves and activity profiles. *arXiv preprint*, 2024.

[24] Dávid Laky, Shuang Xu, J. S. Rodriguez, S. Vaidyaraman, and Salvador García Muñoz. An optimization-based framework to define the probabilistic design space of pharmaceutical processes with model uncertainty. *Processes*, 7(2):96, 2019.

[25] F. Langenbucher. Linearization of dissolution rate curves by the weibull distribution. *Journal of Pharmacy and Pharmacology*, 24(12):979–981, 1972.

[26] J. C. Lee, D. T. Chen, H.-N. Hung, and J. J. Chen. Analysis of drug dissolution data. *Statistics in Medicine*, 18(7):799–814, 1999.

[27] Cheng Li, Jing Wang, Lei Han, and Dong Dong. A simulation model validation method based on functional data analysis. In *Advances in Mechanical Engineering and Industrial Informatics*, pages 516–523. Springer, Berlin, Heidelberg, 2012.

[28] Bo Liu, Chunmei Liu, and Shi Qiao. Development of dissolution curve testing method of coblopasvir hydrochloride capsules. *Capital Food and Drug*, 3:164–168, 2026. In Chinese.

[29] Shaobo Liu, Xiaoyu Cai, Meiyu Shen, and Yi Tsong. In vitro dissolution profile comparison using bootstrap bias-corrected similarity factor f_2. *Journal of Biopharmaceutical Statistics*, 34(1):78–89, 2024.

[30] A. Lourenço, T. Schuster, J. A. Lopes, and A. Kirsch. A non-linear modelling approach to predict the dissolution profile of extended-release tablets. *European Journal of Pharmaceutical Sciences*, 204:106976, 2024.

[31] Mi-Chia Ma, Betty B. C. Wang, Jen-Pei Liu, and Yi Tsong. Assessment of similarity between dissolution profiles. *Journal of Biopharmaceutical Statistics*, 10(2):229–249, 2000.

[32] Wenxia Ma, Chunyan Tang, Qiumin Liu, and Zhongshu Fan. Comparative study on the dissolution curves of febuxostat tablets. *Journal of Pharmaceutical Research*, 36(2):84–88, 2017. In Chinese.

[33] Victor Mangas-Sanjuan, Sarin Colon-Useche, Isabel Gonzalez-Alvarez, Marival Bermejo, and Alfredo Garcia-Arieta. Assessment of the regulatory methods for the comparison of highly variable dissolution profiles. *The AAPS Journal*, 18(6):1550–1561, 2016.

[34] Vicente Mangas-Sanjuán, Sebastián Colón-Useche, Isabel González-Álvarez, María Bermejo, and Amparo García-Arieta. Assessment of the regulatory methods for the comparison of highly variable dissolution profiles. *The AAPS Journal*, 18(6):1550–1561, 2016.

[35] Marilyn N. Martinez and Xiongce Zhao. A simple approach for comparing the *In Vitro* dissolution profiles of highly variable drug products: a proposal. *The AAPS Journal*, 20(4):78, 2018.

[36] John W. Mauger, Daniel Chilko, and Stephen Howard. On the analysis of dissolution data. *Drug Development and Industrial Pharmacy*, 12(7):969–992, 1986.

[37] M. D. McKay, R. J. Beckman, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245, 1979.

[38] Jonathan P. McMullen, Brian M. Wyvratt, Cynthia M. Hong, and Akasha K. Purohit. Integrating functional principal component analysis with data-rich experimentation for enhanced drug substance development. *Organic Process Research & Development*, 28(3):719–728, 2024.

[39] A. Mendyk, R. Jachowicz, M. Kurek, P. Dorozynski, Z. Pedzich, and J. Szlek. Kinetds: An open source software for dissolution test data analysis. *Dissolution Technologies*, 19(1):6–11, 2012.

[40] Andrzej Mendyk, Roman Jachowicz, Krzysztof Fijorek, Przemysław Dorożyński, Piotr Kulinowski, and Stanisław Polak. Open-source software for the simulation of f_2 distribution in cases of large variability in dissolution profiles. *Dissolution Technologies*, 19(2):20–25, 2012.

[41] Ivana Mitrevska, Ljupco Pejov, Marija Jovanovska, Suzan Memed-Sejfulah, Katerina Brezovska, et al. Conventional and multivariate statistical methods for evaluation of in vitro dissolution similarity of bisoprolol film-coated tablets. *International Journal of Pharmacy and Chemistry*, 6(2):16–25, 2020.

[42] J. W. Moore and H. H. Flanner. Mathematical comparison of dissolution profiles. *Pharmaceutical Technology*, 20(6):64–74, 1996.

[43] M. Mrad et al. Spectroscopy-based partial prediction of in vitro dissolution profile using artificial neural networks. *Periodica Polytechnica Electrical Engineering and Computer Science*, 2022.

[44] Jan Muselík, Alena Komersová, Kateřina Kubová, Kevin Matzick, and Barbora Skalická. A critical overview of FDA and EMA statistical methods to compare *In Vitro* drug dissolution profiles of pharmaceutical products. *Pharmaceutics*, 13(10):1703, 2021.

[45] Steven Novick, Yan Shen, Harry Yang, John Peterson, Dave LeBlond, and Stan Altan. Dissolution curve comparisons through the f_2 parameter: A bayesian extension of the f_2 statistic. *Journal of Biopharmaceutical Statistics*, 25(2):351–371, 2015.

[46] Tadej Ojsteršek, Franc Vrečer, and Grega Hudovornik. Comparative fitting of mathematical models to carvedilol release profiles obtained from hypromellose matrix tablets. *Pharmaceutics*, 16(4):498, 2024.

[47] K. K. Peh et al. Use of artificial neural networks to predict drug dissolution profiles and evaluation of network performance using similarity factor. *Pharmaceutical Research*, 2000.

[48] D. A. Piscitelli and D. Young. Setting dissolution specifications for modified-release dosage forms. *Advances in Experimental Medicine and Biology*, 423:159–166, 1997.

[49] Tony Pourmohamad, Mehdi S. Mir, and Vinod P. Shah. Statistical modeling approaches for the comparison of dissolution profiles. *AAPS PharmSciTech*, 23(2):1–12, 2022.

[50] James O. Ramsay and Bernard W. Silverman. *Functional Data Analysis*. Springer, New York, 2 edition, 2005.

[51] A. Ruiz-Picazo, I. González-Álvarez, M. Bermejo, and M. González-Álvarez. New mathematical model from dynamic dissolution rate tests. *European Journal of Pharmaceutical Sciences*, page 106864, 2024.

[52] S. Sadray et al. Dissolution profile comparison: Model dependent and model independent approaches. *International Journal of Pharma and Bio Sciences*, 2010.

[53] H. Saranadasa. Defining similarity of dissolution profiles: Through hotelling's t2 statistic. *Pharmaceutical Technology*, 2001.

[54] George A. F. Seber and C. J. Wild. *Nonlinear Regression*. Wiley, 2003.

[55] Vinod P. Shah, Yi Tsong, Pradeep Sathe, and Jen-Fu Liu. In-vitro dissolution profile comparison—statistics and analysis, model-dependent approach. *Pharmaceutical Research*, 13:1799–1803, 1996.

[56] Deniz Sigirli and Ilker Ercan. Examining growth with statistical shape analysis and comparison of growth models. *Journal of Modern Applied Statistical Methods*, 11(2):19, 2012.

[57] B. W. Silverman. Incorporating parametric effects into functional principal components analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(4):673–689, 1995.

[58] Ronald D. Snee. A strategy for the analysis of dissolution profiles. *Pharmaceutical Engineering*, 2019.

[59] M. Sousa et al. A comparative study of two data-driven modeling approaches to predict drug release from extended-release matrix tablets. *International Journal of Pharmaceutics*, 2025. Functional data analysis and FPCA compared with artificial neural networks for dissolution prediction.

[60] Y. Tsong, P. M. Sathe, and V. P. Shah. In vitro dissolution profile comparison. In S.-C. Chow, editor, *Encyclopedia of Biopharmaceutical Statistics*, pages 428–432. Marcel Dekker, New York, 2003.

[61] Yi Tsong, Thomas Hammerstrom, Pradeep Sathe, and Vinod P. Shah. Statistical assessment of mean differences between two dissolution data sets. *Drug Information Journal*, 30:1105–1112, 1996.

[62] U.S. Food and Drug Administration. Quality by design for andas: An example for immediate-release dosage forms, 2012. Guidance for Industry.

[63] U.S. Food and Drug Administration. Dissolution testing of immediate release solid oral dosage forms, 2019. Guidance for Industry.

[64] R. Yokoyama, G. Kimura, C. M. Schlepütz, J. Huwyler, and M. Puchkov. Investigation of disintegration and dissolution behavior of mefenamic acid drug formulation using numeric solution of noyes-whitney equation with cellular automata model on microtomographic surfaces and rational arrangements of tablet components. *Preprints*, 2018.

[65] H. Yoshida et al. Comparison of dissolution similarity assessment methods for products with large variations: f2 statistics and model-independent multivariate confidence region procedure for dissolution profiles of multiple oral products. *Biological & Pharmaceutical Bulletin*, 2017.

[66] G. Udny Yule. On a method of investigating periodicities in disturbed series, with special reference to wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society A*, 226:267–298, 1927.

[67] Shuyan Zhai, Thomas Mathew, and Yi Huang. Comparison of drug dissolution profiles: a proposal based on tolerance limits. *Statistics in Medicine*, 35(29):5464–5476, 2016.

Appendix A

Supplementary Mathematical Details for Smoothing Methods

This appendix provides additional mathematical and implementation-level details for the smoothing approaches introduced in Chapter 3. In the main text, emphasis is placed on conceptual definitions and modeling objectives, while technical details (basis construction, penalty matrices, and constraint formulations) are deferred here.

A.1 Notation

Let $y_{ij}(t_k)$ denote the observed dissolution percentage for tablet j in batch i at time point t_k, and let the batch-level mean profile be

$$y_i(t_k) \;=\; \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}(t_k),$$

where n_i is the number of tablets in batch i. Smoothing methods estimate a continuous function $x_i(t)$ for each batch.

For a set of observation times $\{t_k\}_{k=1}^{m}$, define $\mathbf{y}_i = (y_i(t_1), \ldots, y_i(t_m))^\top$.

A.1.1 Basis Dimension in Functional Data Analysis

In functional data analysis, a smooth function $x(t)$ is commonly represented using a finite-dimensional basis expansion,

$$x(t) = \sum_{k=1}^{K} c_k\, \psi_k(t),$$

where K denotes the number of basis functions. As discussed by Ramsay and Silverman, the choice of K is not arbitrary but is linked to the resolution of the observed data, the smoothness of the underlying process, and the order of the basis functions.

For spline-based representations, the basis dimension is implicitly determined by the spline order and the underlying knot sequence defined over the domain of t. In particular, the number of basis functions increases with the number of knots and the spline order, but individual knots are not treated as model parameters in functional regression.

In practice, FDA emphasizes that the effective complexity of the fitted function is governed primarily by the smoothing or roughness penalty rather than the raw basis dimension. As a result, it is common to select a basis dimension that is sufficiently large to capture potential features of the data and to rely on penalization to control overfitting.

Following this principle, the analyses in this thesis use spline bases with fixed basis dimensions chosen to accommodate the observed sampling density, while smoothness is regulated through penalty parameters selected by data-driven criteria. This approach aligns with standard FDA practice and ensures stable functional representations across batches.

As noted in Ramsay and Silverman [50], for spline bases the number of basis functions K is determined implicitly by the chosen spline order and knot sequence over the domain. In functional data analysis, these construction details are treated as part of the basis definition rather than as parameters to be estimated. Consequently, FDA practice emphasizes selecting a basis of sufficient dimension and controlling effective model complexity through smoothing penalties.

A.2 B-splines: Basis Construction and Design Matrix

B-spline smoothing represents each profile as

$$x_i(t) = \sum_{k=1}^{K} c_{ik}\, B_k(t),$$

where $\{B_k(t)\}_{k=1}^{K}$ are B-spline basis functions and $\{c_{ik}\}$ are the corresponding coefficients. Here, K denotes the dimension of the spline basis, i.e., the number of basis functions, and does not correspond directly to the number of knots. Knot placement and spline order are implicit in the construction of the basis functions and are not treated as explicit model parameters.

Let $\mathbf{B}$ denote the design matrix evaluated at observation times:

$$\mathbf{B} = \begin{pmatrix} B_1(t_1) & \cdots & B_K(t_1) \\ \vdots & \ddots & \vdots \\ B_1(t_m) & \cdots & B_K(t_m) \end{pmatrix}.$$

Then the fitted values at observed time points are $\widehat{\mathbf{y}}_i = \mathbf{B}\mathbf{c}_i$, where $\mathbf{c}_i = (c_{i1}, \ldots, c_{iK})^\top$. Unpenalized least squares estimation solves

$$\min_{\mathbf{c}_i} (\mathbf{y}_i - \mathbf{B}\mathbf{c}_i)^\top (\mathbf{y}_i - \mathbf{B}\mathbf{c}_i).$$

Details such as spline order and knot placement may be selected based on considerations (e.g., sparse sampling) and numerical stability.

A.3 P-splines: Coefficient and Derivative Penalty Formulations

In this thesis, penalized spline (P-spline) smoothing is implemented using B-spline bases augmented with roughness penalties. Let

$$x_i(t) = \sum_{k=1}^{K} c_{ik}\, B_k(t)$$

denote the spline representation for batch i, where $\{B_k(t)\}$ are B-spline basis functions and $\mathbf{c}_i = (c_{i1}, \ldots, c_{iK})^\top$ are the corresponding coefficients.

Two standard forms of roughness control are considered: penalties on the curvature of the fitted function and penalties on differences between adjacent spline coefficients.

A.3.1 Second-Derivative (Curvature) Penalty

A commonly used roughness penalty in spline smoothing penalizes the integrated squared second derivative of the fitted curve,

$$\lambda \int \{x_i''(t)\}^2 \, dt,$$

where $\lambda \geq 0$ controls the trade-off between fidelity to the data and smoothness of the estimated function. This formulation directly discourages rapid changes in curvature and is widely used in functional data analysis when the underlying function is sufficiently well-sampled.

In datasets with dense time grids or simulated observations, this curvature-based penalty was successfully applied to obtain smooth and stable functional representations.

A.3.2 Coefficient-Based Difference Penalty

Alternatively, smoothness can be enforced directly on the spline coefficients through finite-difference penalties. In this formulation, the penalized least squares problem is

$$\min_{\mathbf{c}_i} \left[(\mathbf{y}_i - \mathbf{B}\mathbf{c}_i)^\top (\mathbf{y}_i - \mathbf{B}\mathbf{c}_i) + \lambda \, \mathbf{c}_i^\top \mathbf{P} \mathbf{c}_i \right],$$

where $\mathbf{B}$ is the spline design matrix and $\mathbf{P}$ is a penalty matrix derived from a difference operator.

Let $\mathbf{D}_d$ denote the dth-order difference matrix (commonly $d = 2$). Then

$$\mathbf{c}_i^\top \mathbf{P} \mathbf{c}_i = \|\mathbf{D}_d \mathbf{c}_i\|^2 = \sum_k (\Delta^d c_{ik})^2, \quad \text{with} \quad \mathbf{P} = \mathbf{D}_d^\top \mathbf{D}_d.$$

This penalty discourages abrupt changes in adjacent spline coefficients and yields smooth fitted curves while maintaining numerical stability.

The minimizer has the closed-form solution

$$\widehat{\mathbf{c}}_i = (\mathbf{B}^\top \mathbf{B} + \lambda \mathbf{P})^{-1} \mathbf{B}^\top \mathbf{y}_i,$$

provided the matrix is invertible.

A.3.3 Considerations

In the dataset B analyzed in this thesis, the coefficient-based penalty was adopted in practice. The curvature-based second-derivative penalty led to numerical or implementation issues under sparse sampling and the chosen spline representation in R. Penalizing coefficient differences provided a stable and reproducible smoothing procedure across batches.

For other datasets with denser sampling or simulated designs, the second-derivative penalty was feasible and produced comparable smooth functional representations. In all cases, the smoothing parameter λ was selected using data-driven criteria.

A.4 Monotone P-splines (Exploratory)

Monotone P-spline variants impose constraints intended to ensure $x_i(t)$ is non-decreasing. One approach enforces monotonicity by constraining first differences of coefficients to be nonnegative:

$$\Delta c_{ik} \geq 0 \quad \text{for all relevant } k,$$

which can be written in matrix form as

$$\mathbf{D}_1 \mathbf{c}_i \geq \mathbf{0}.$$

In this setting, estimation becomes a constrained optimization problem:

$$\min_{\mathbf{c}_i} \left[(\mathbf{y}_i - \mathbf{B}\mathbf{c}_i)^\top (\mathbf{y}_i - \mathbf{B}\mathbf{c}_i) + \lambda \, \mathbf{c}_i^\top \mathbf{P} \mathbf{c}_i \right] \quad \text{subject to } \mathbf{D}_1 \mathbf{c}_i \geq \mathbf{0}.$$

These approaches were explored to eliminate non-physical decreases but were found to be sensitive to sparse sampling and tuning choices and were not adopted as the primary smoothing strategy.

A.5 I-splines and Penalized I-spline Variants (Exploratory)

I-splines provide a monotone basis construction derived from M-spline basis functions. M-splines are nonnegative, piecewise polynomial basis functions defined over a knot sequence and serve as the building blocks for monotone spline representations. Because M-splines are everywhere nonnegative, integrating them yields basis functions that are non-decreasing over the domain, providing a direct mechanism for enforcing monotonicity.

A.5.1 M-spline Basis Functions

Let $\{\kappa_1, \ldots, \kappa_{K+p+1}\}$ denote a non-decreasing knot sequence over the domain of t, where p is the polynomial order. The kth M-spline basis function of order p is defined recursively as

$$M_{k,p}(t) = \frac{p+1}{\kappa_{k+p+1} - \kappa_k} \left[(t - \kappa_k) M_{k,p-1}(t) + (\kappa_{k+p+1} - t) M_{k+1,p-1}(t) \right],$$

with zeroth-order M-splines given by

$$M_{k,0}(t) = \begin{cases} \dfrac{1}{\kappa_{k+1} - \kappa_k}, & \kappa_k \leq t < \kappa_{k+1}, \\ 0, & \text{otherwise.} \end{cases}$$

M-spline basis functions are nonnegative and integrate to one over their support. These properties make them suitable for constructing monotone function bases through integration.

A.5.2 I-spline Basis Construction

I-spline basis functions are defined as integrals of M-spline basis functions. Specifically, the kth I-spline basis function is given by

$$I_k(t) = \int_{\kappa_1}^{t} M_{k,p}(u)\, du.$$

Because M-splines are nonnegative, each I-spline basis function is non-decreasing over the domain of t. As a result, any nonnegative linear combination of I-splines yields a monotone function.

Using I-spline basis functions $\{I_k(t)\}$, the fitted dissolution curve for batch i is represented as

$$x_i(t) = \sum_{k=1}^{K} \alpha_{ik}\, I_k(t), \qquad \alpha_{ik} \geq 0,$$

where nonnegativity of the coefficients $\{\alpha_{ik}\}$ guarantees that $x_i(t)$ is a non-decreasing function of time.

Let $\mathbf{I}$ denote the design matrix with entries $I_k(t_j)$ evaluated at the observed time points. Coefficients are estimated by solving the nonnegative least squares problem

$$\min_{\boldsymbol{\alpha}_i} (\mathbf{y}_i - \mathbf{I}\boldsymbol{\alpha}_i)^\top (\mathbf{y}_i - \mathbf{I}\boldsymbol{\alpha}_i) \quad \text{subject to } \boldsymbol{\alpha}_i \geq \mathbf{0}.$$

A.5.3 Anchoring at the First Observed Time Point

In practice, dissolution experiments rarely include observations at time $t = 0$, and measurements typically begin at a later time point $t_1 > 0$. Although cumulative dissolution is physically zero at the origin, enforcing the constraint $x_i(0) = 0$ would require extrapolation outside the observed data range and may distort the fitted curve near early time regions.

To avoid this issue, I-spline models were anchored at the first observed time point by enforcing

$$x_i(t_1) = y_i(t_1),$$

where t_1 denotes the earliest sampling time and $y_i(t_1)$ is the observed batch-level mean dissolution at that time.

Anchoring at t_1 ensures that the fitted curve passes exactly through the first observed data point, preserving fidelity to the experimental measurements while avoiding extrapolation-based artifacts. This constraint is particularly important for monotone spline methods, which may otherwise introduce artificial offsets at early time points due to the absence of data near the origin.

Implementation of this anchoring was achieved either by fixing the corresponding linear constraint during coefficient estimation or by adjusting the basis representation so that the fitted curve is conditioned on the first observation.

A.6 Effect of Anchoring on Monotone Spline Fits

To assess the impact of anchoring choices on monotone spline smoothing, additional diagnostic analyses were conducted in which I-spline and penalized I-spline models were fitted without enforcing an anchor at the first observed time point.

In the absence of anchoring, the fitted curves may exhibit artificial offsets near the initial time region or fail to align with the first observed dissolution measurement. This issue is particularly evident when early sampling times are sparse, where the spline basis lacks sufficient constraint to accurately capture the initial release behavior. As a result, the fitted trajectory may deviate from the expected physical behavior of dissolution processes, which typically start near zero.

Imposing an anchoring condition at the first observed time point enforces agreement between the fitted curve and the observed data at t_1, thereby stabilizing the fit in the early time region. This constraint reduces boundary artifacts and leads to more physically interpretable monotone dissolution profiles.

Based on these observations, anchored monotone spline models were adopted as the final smoothing strategy throughout this thesis.

A.6.1 Penalized I-spline Variants

While unpenalized I-spline smoothing guarantees monotonicity, additional regularization can be introduced to control smoothness and reduce sensitivity to sampling variability. In this work, two penalized I-spline formulations were explored: coefficient-based difference penalties and curvature-based second-derivative penalties.

Coefficient-Based Difference Penalty

Using the I-spline representation

$$x_i(t) = \sum_{k=1}^{K} \alpha_{ik}\, I_k(t), \qquad \alpha_{ik} \geq 0,$$

smoothness can be enforced by penalizing finite differences of the coefficient sequence. The penalized optimization problem is

$$\min_{\boldsymbol{\alpha}_i} \left[\|\mathbf{y}_i - \mathbf{I}\boldsymbol{\alpha}_i\|^2 + \lambda\, \boldsymbol{\alpha}_i^\top \mathbf{Q} \boldsymbol{\alpha}_i \right] \quad \text{subject to } \boldsymbol{\alpha}_i \geq \mathbf{0},$$

where $\mathbf{Q} = \mathbf{D}_d^\top \mathbf{D}_d$ is constructed from a dth-order finite difference matrix $\mathbf{D}_d$ (typically $d = 2$). This penalty discourages abrupt changes between adjacent I-spline coefficients while preserving monotonicity.

Coefficient-based penalization was numerically stable under sparse sampling and aligns naturally with P-spline regularization strategies while maintaining the monotonicity constraints imposed by the I-spline basis.

Second-Derivative (Curvature) Penalty

Alternatively, smoothness can be imposed directly on the fitted function through a curvature-based penalty on the second derivative:

$$\min_{\boldsymbol{\alpha}_i} \left[\|\mathbf{y}_i - \mathbf{I}\boldsymbol{\alpha}_i\|^2 + \lambda \int \{x_i''(t)\}^2 \, dt \right] \quad \text{subject to } \boldsymbol{\alpha}_i \geq \mathbf{0}.$$

This formulation directly penalizes local curvature in the estimated dissolution curve and is commonly used in functional data analysis when the underlying function is sufficiently well-sampled.

In practice, curvature-based penalization was feasible in simulation studies and datasets with dense time grids. However, under sparse sampling or certain software implementations, numerical instability was observed, motivating the use of coefficient-based penalties in those settings.

Summary of Penalized I-spline Exploration

Both penalization strategies preserve the monotonicity guaranteed by the I-spline basis while providing additional control over smoothness. Despite producing physically plausible monotone curves, penalized I-spline methods were sensitive to penalty selection and sampling density. Consequently, they were investigated as exploratory alternatives and were not adopted as the primary smoothing strategy in the final analysis.

Appendix B

Additional Plots

This appendix contains supplementary plots that are not included in the main text due to space and readability considerations. These include full reconstruction plots and optimization plots for Datasets A and B under both full-model and variable-selection settings.

B.1 Reconstruction Plots

Under the full-model specification, the standard and anchored parametric pipelines may yield identical fitted curves when the anchoring effect is absorbed by the regression structure. Accordingly, when the resulting reconstruction plots are identical, only one set of curves is presented.

B.1.1 Dataset A

Full-model pipelines

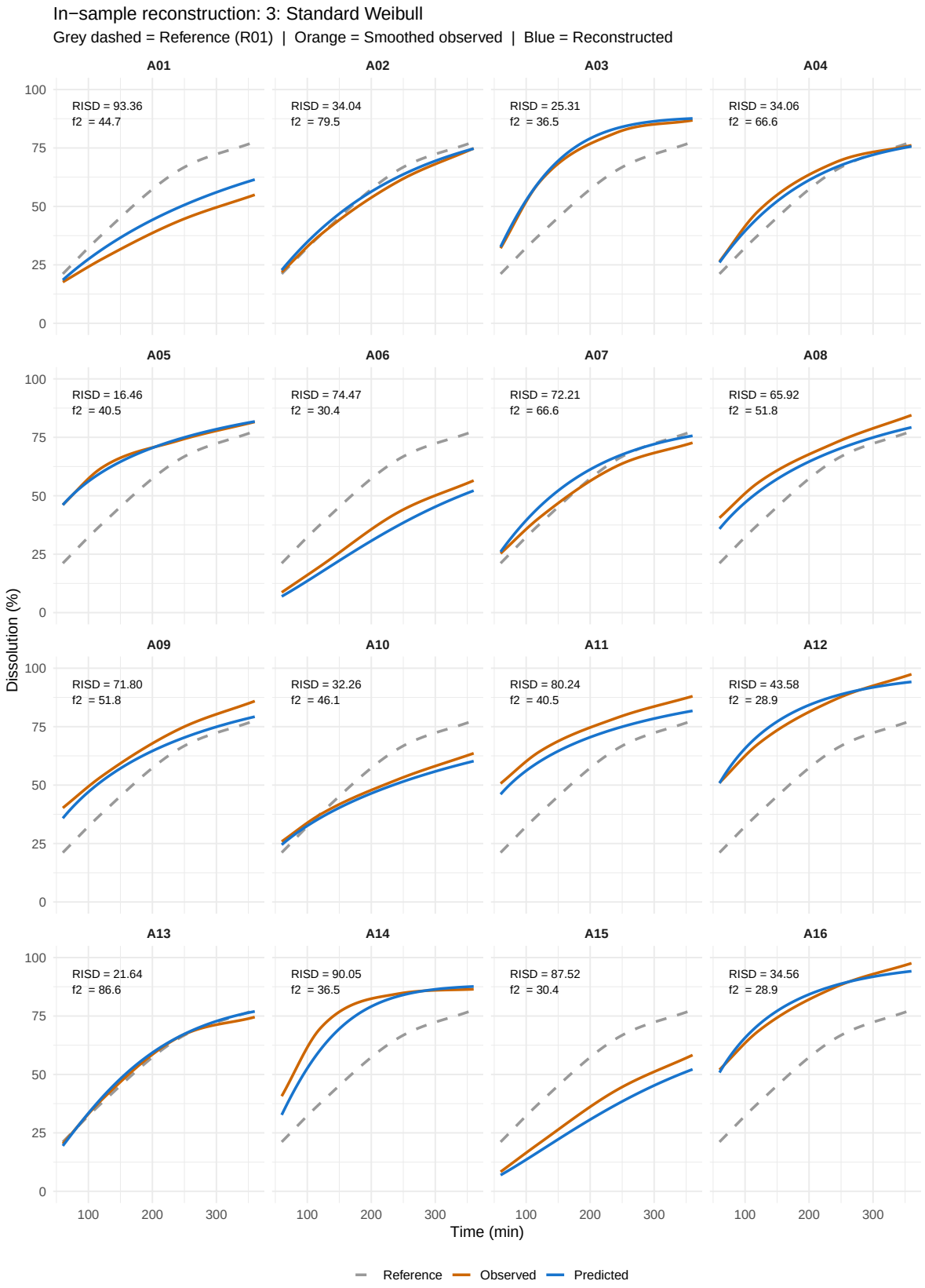


Figure B.1: Full set of reconstructed dissolution curves for Dataset A under the Weibull model. The standard and anchored formulations produce identical results under the full-model specification.

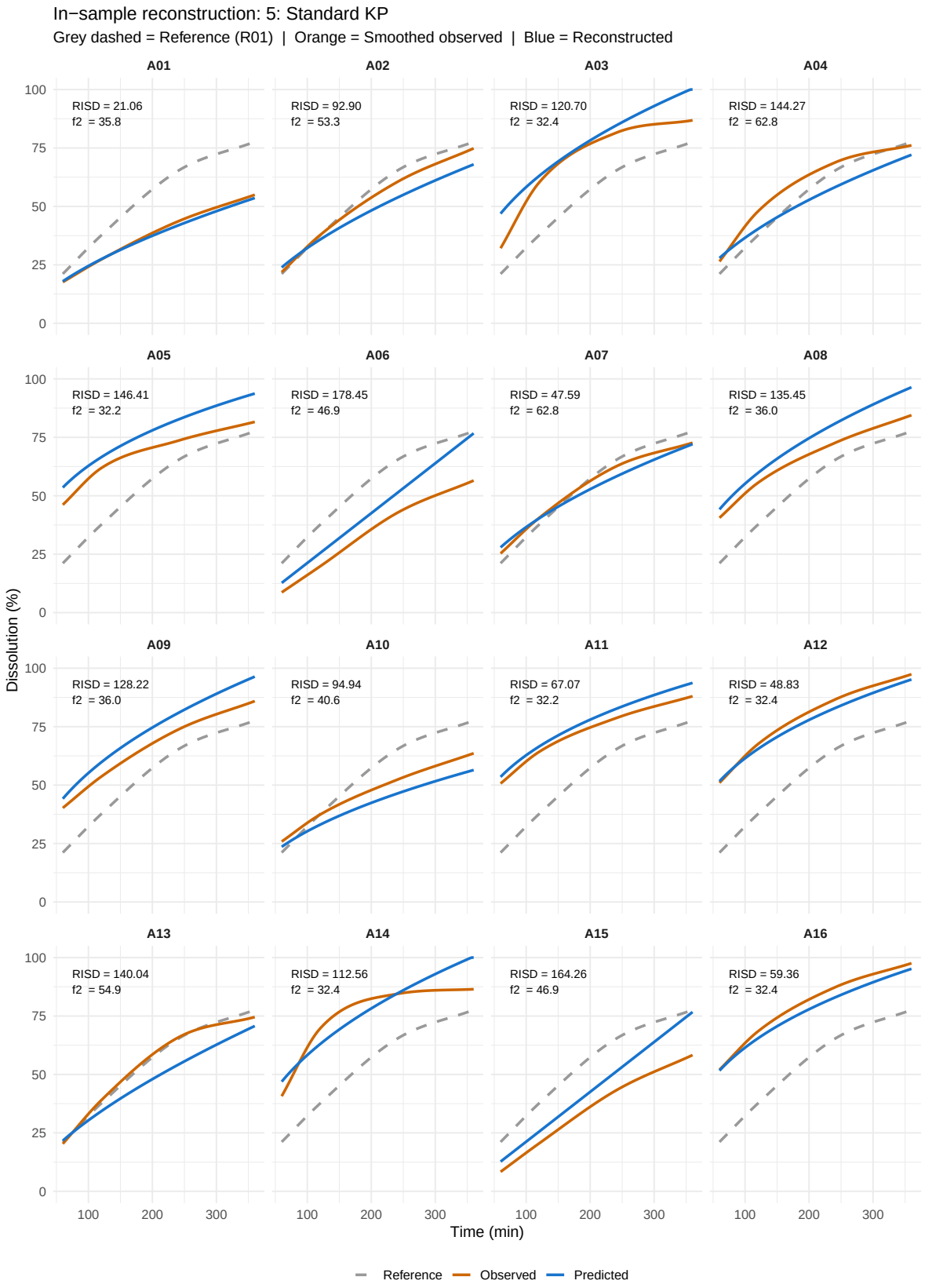


Figure B.2: Full set of reconstructed dissolution curves for Dataset A under the KP model. The standard and anchored formulations produce identical results under the full-model specification.

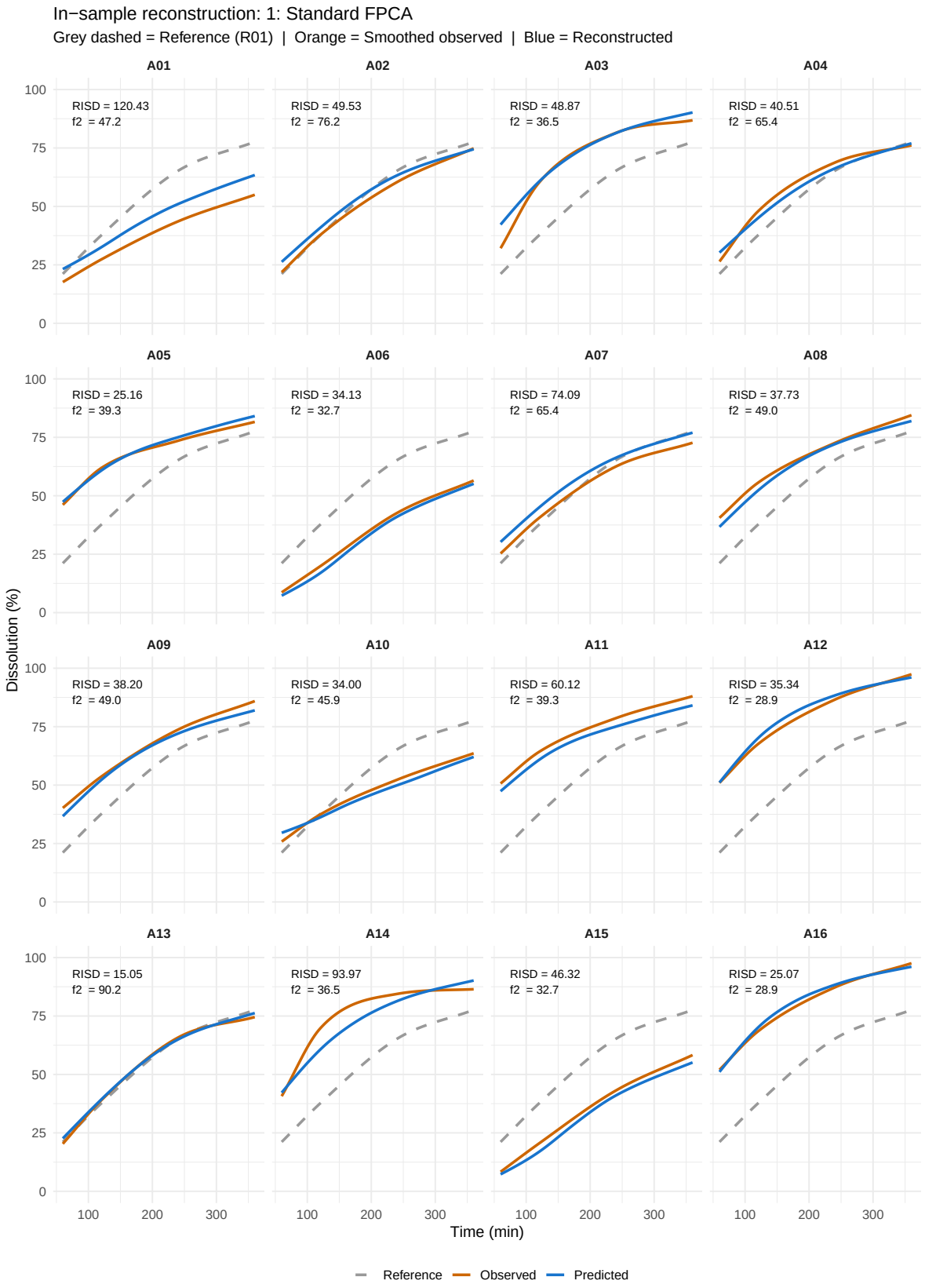


Figure B.3: Full set of reconstructed dissolution curves for Dataset A under the standard FPCA model.

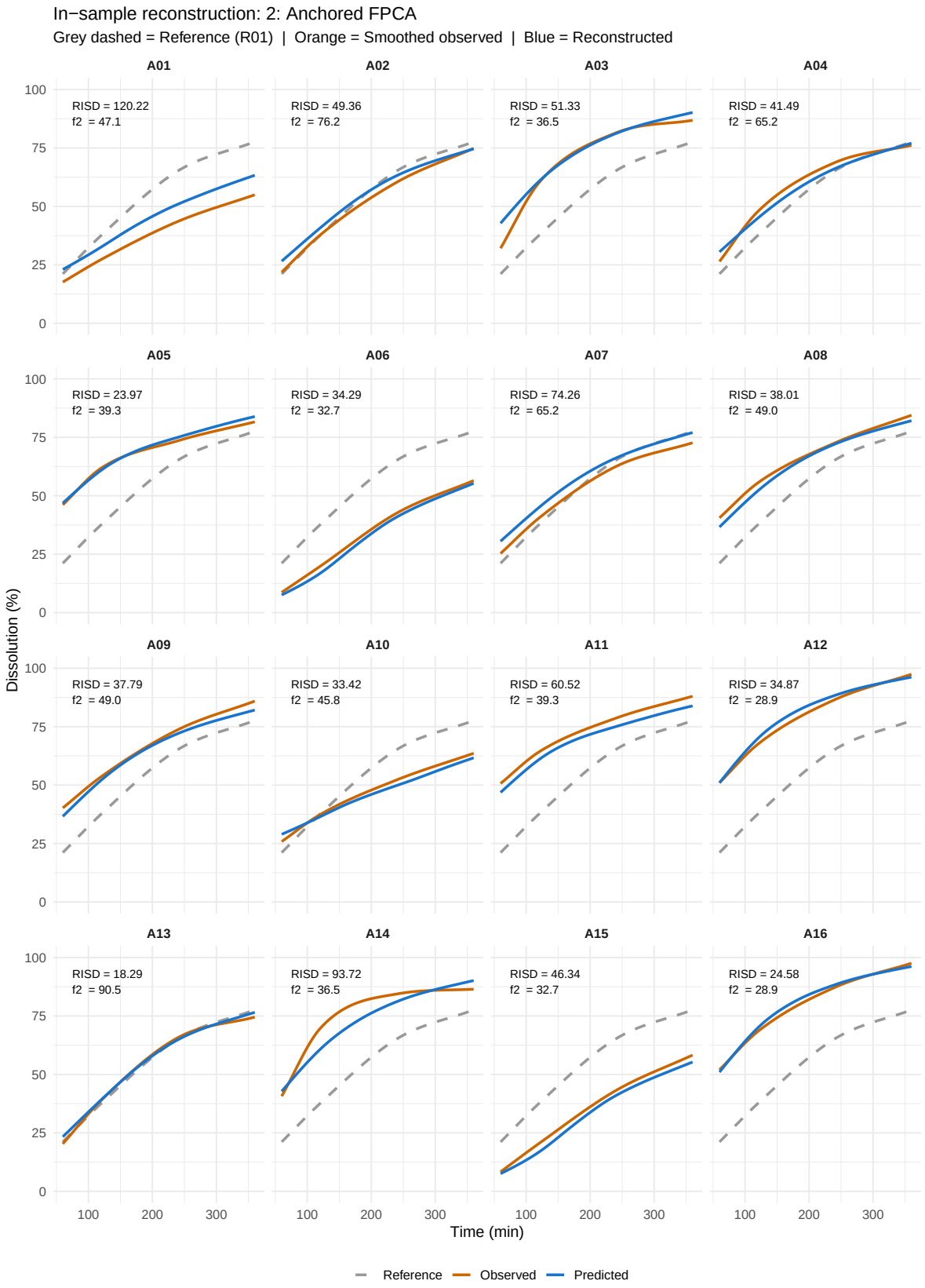


Figure B.4: Full set of reconstructed dissolution curves for Dataset A under the anchored FPCA model.

Variable-selection pipelines

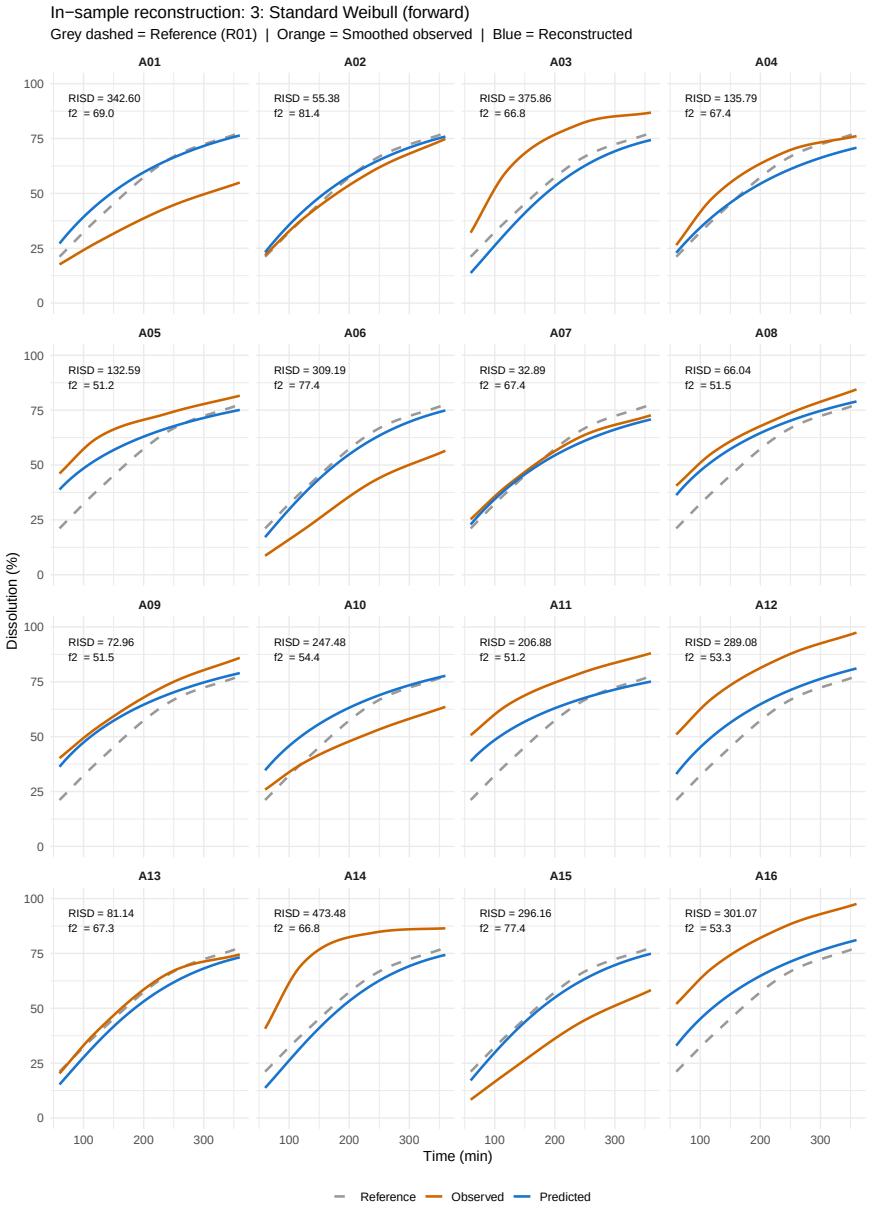


Figure B.5: Reconstructed dissolution curves for Dataset A under the standard Weibull model with variable selection.

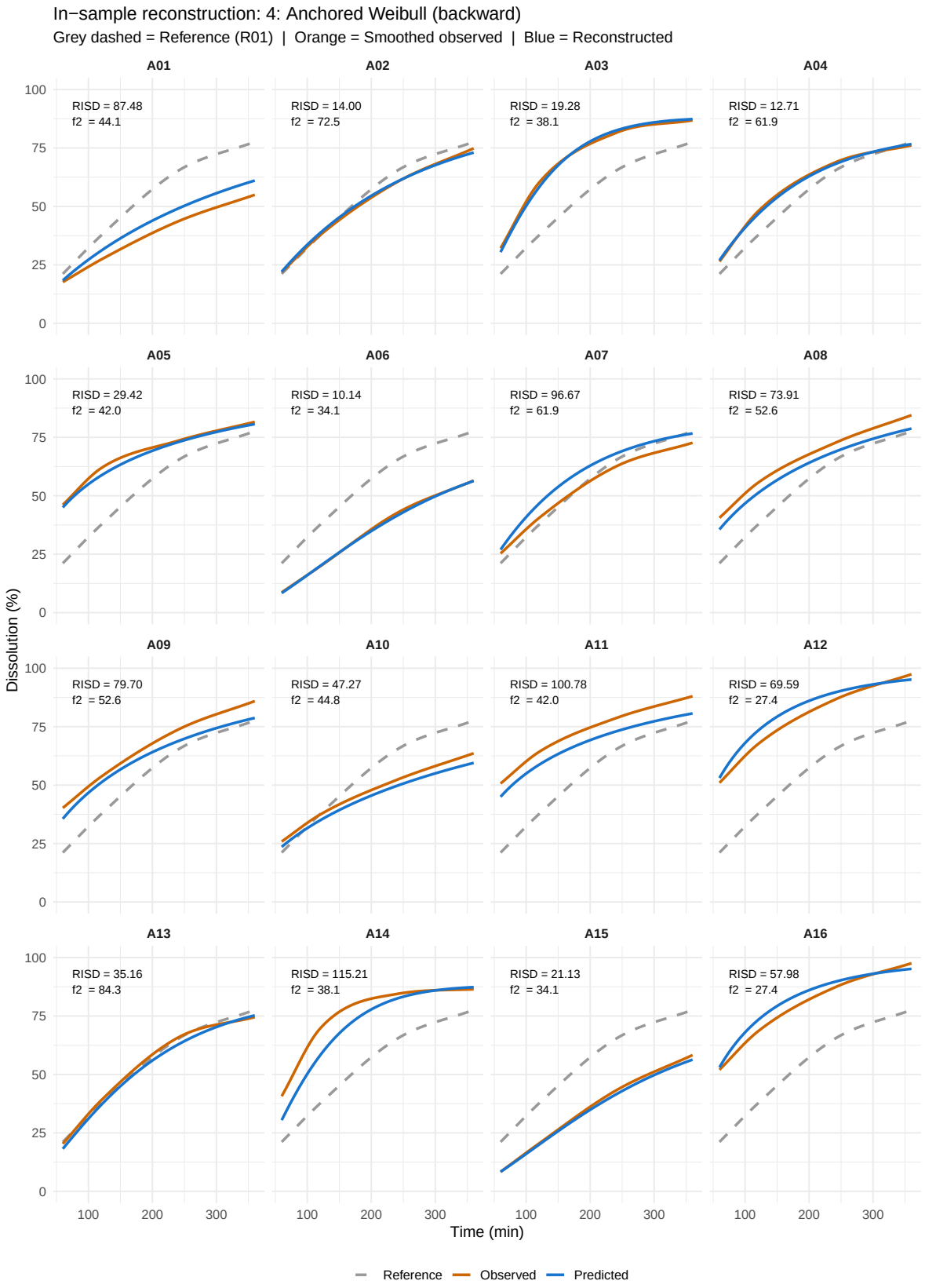


Figure B.6: Reconstructed dissolution curves for Dataset A under the anchored Weibull model with variable selection.

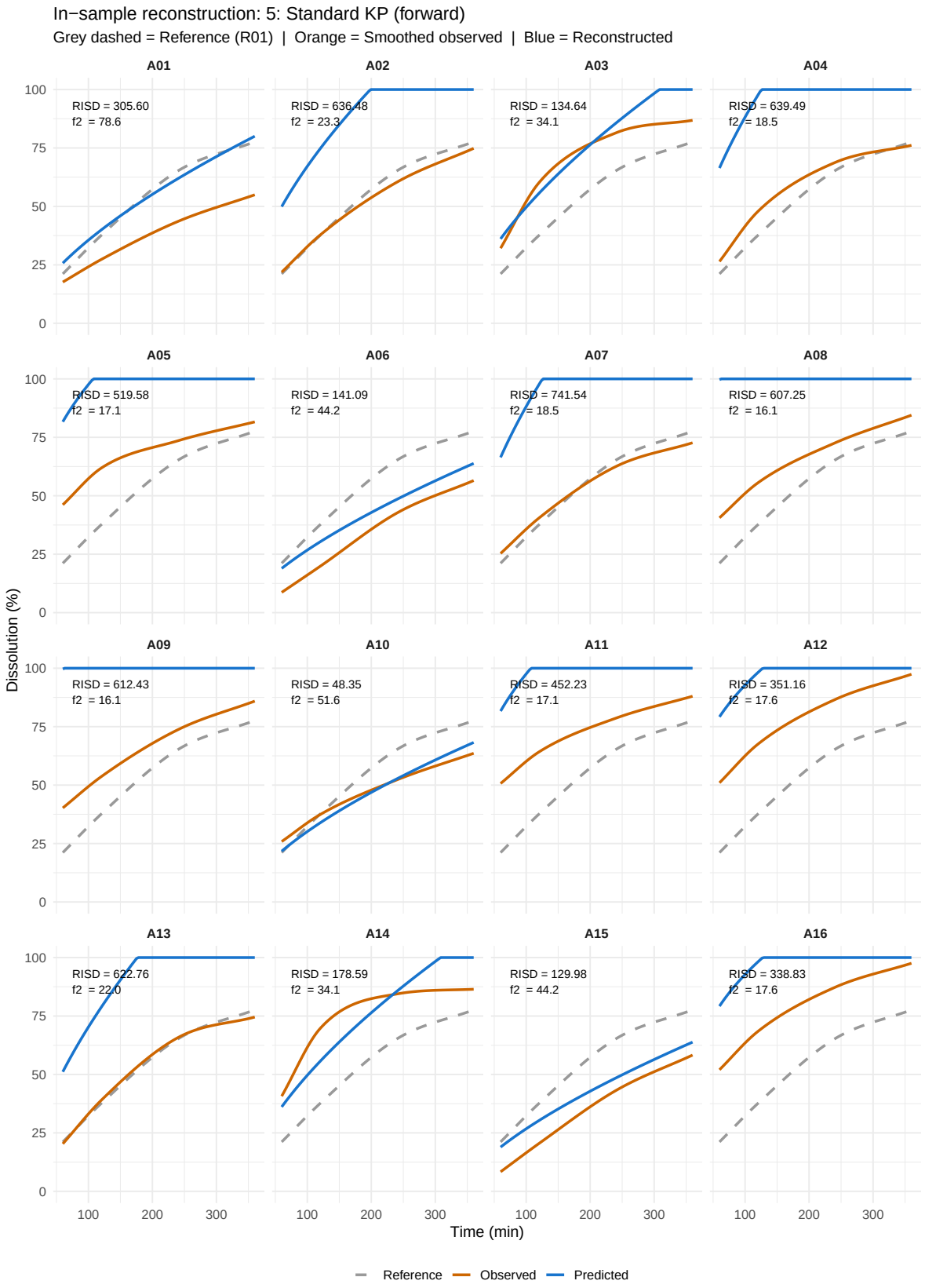


Figure B.7: Reconstructed dissolution curves for Dataset A under the standard KP model with variable selection.

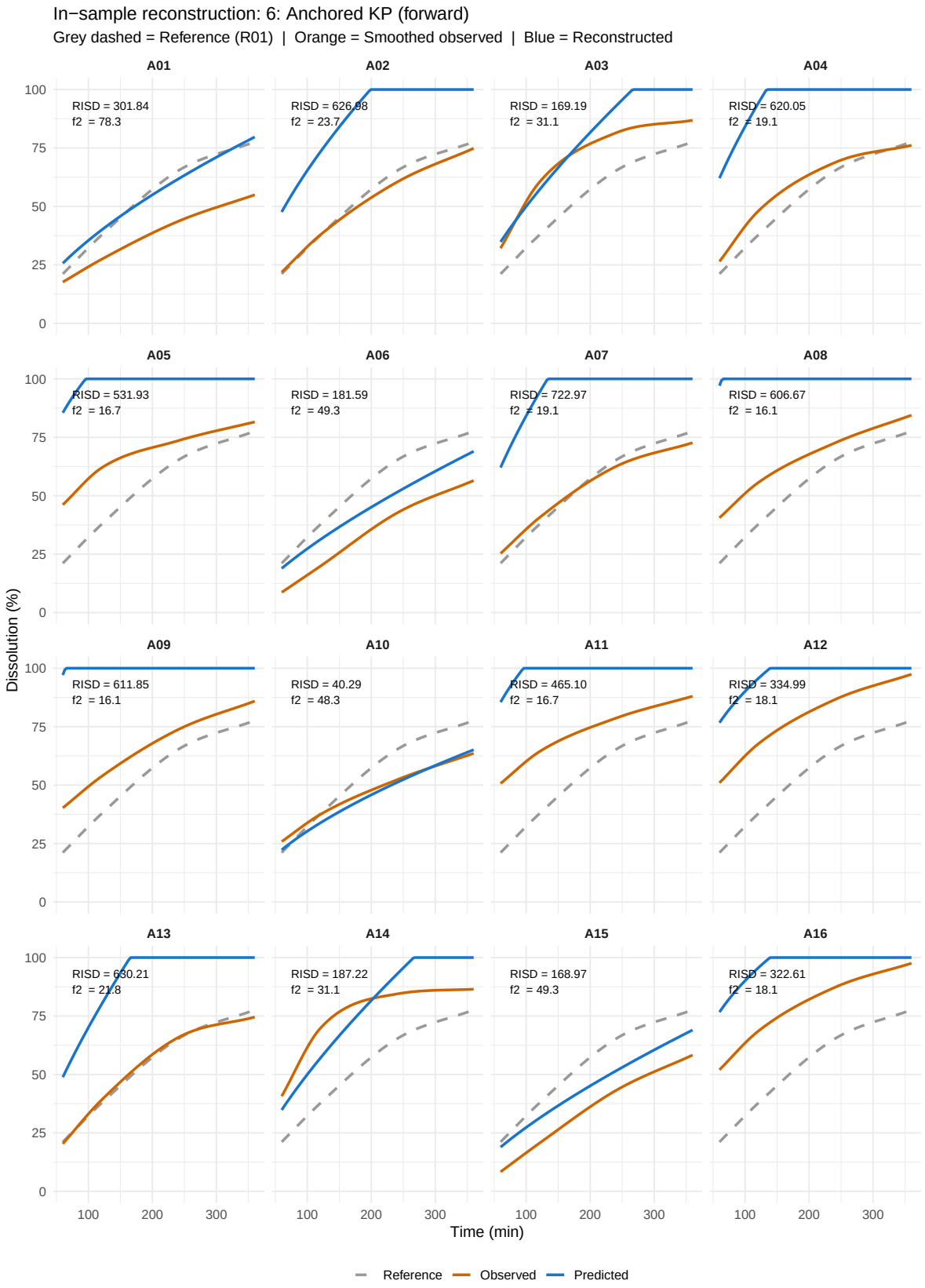


Figure B.8: Reconstructed dissolution curves for Dataset A under the anchored KP model with variable selection.

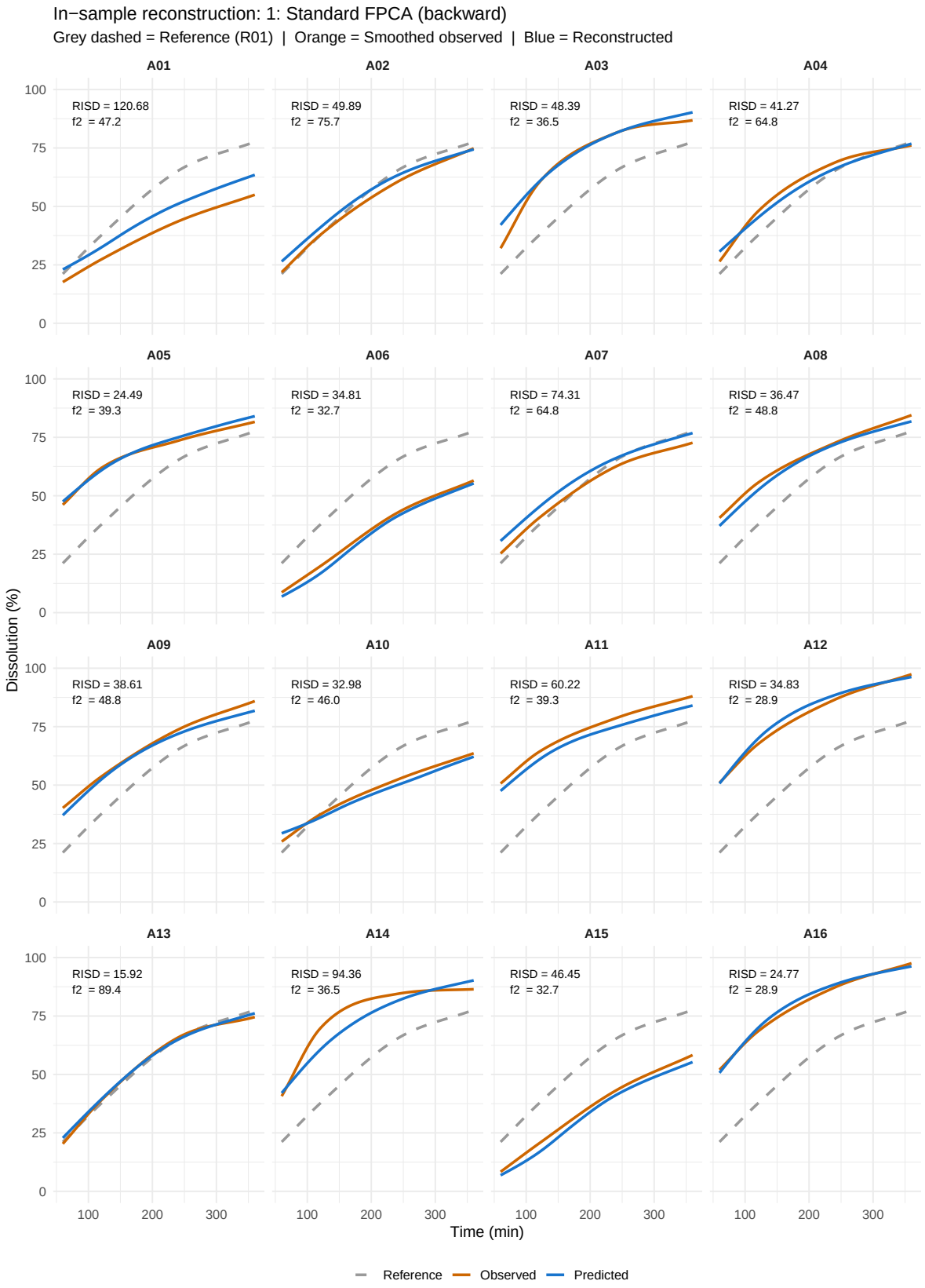


Figure B.9: Reconstructed dissolution curves for Dataset A under the standard FPCA model with variable selection.

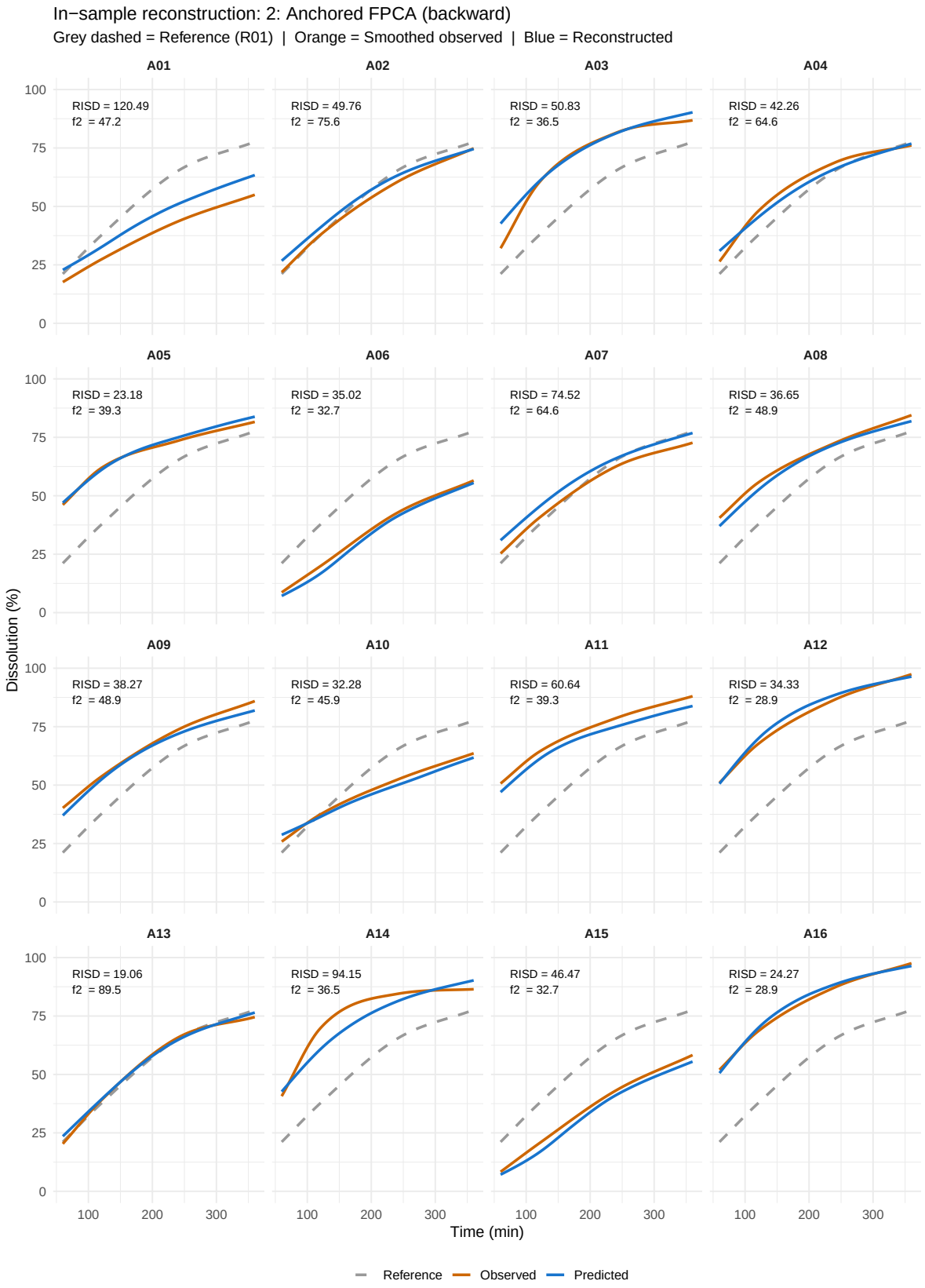


Figure B.10: Reconstructed dissolution curves for Dataset A under the anchored FPCA model with variable selection.

B.1.2 Dataset B

Full-model pipelines

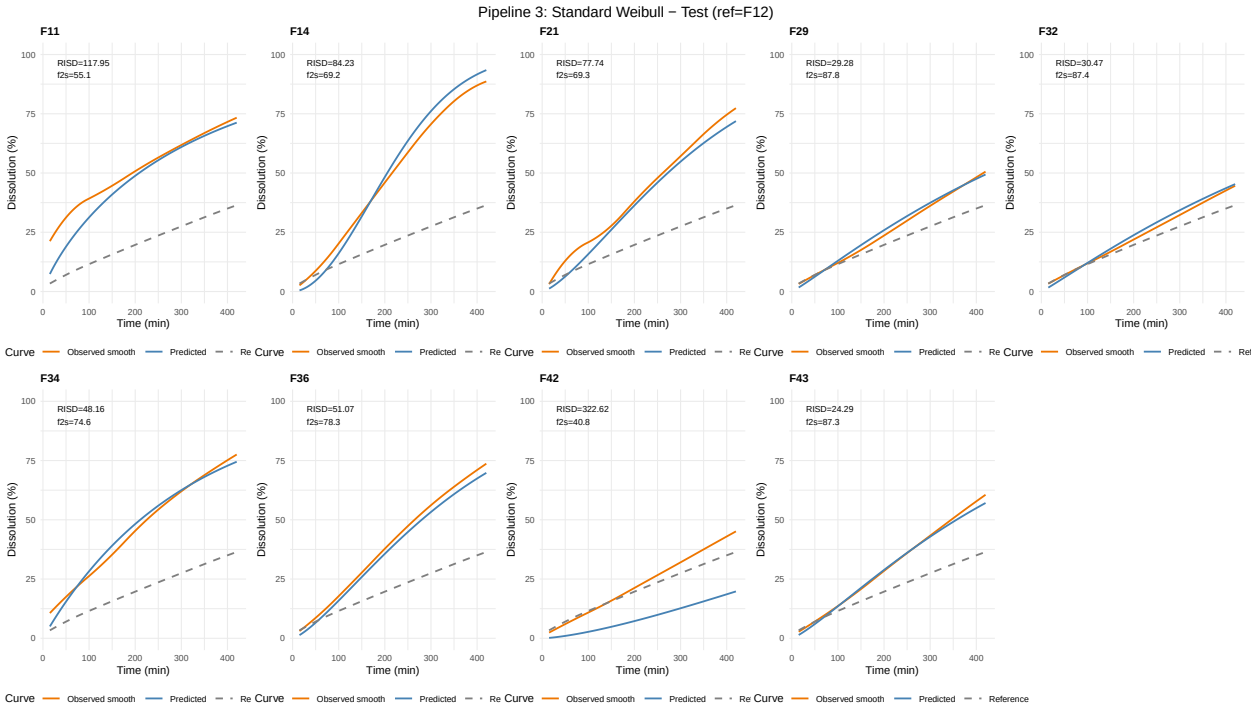


Figure B.11: Full set of reconstructed dissolution curves for Dataset B under the Weibull model.

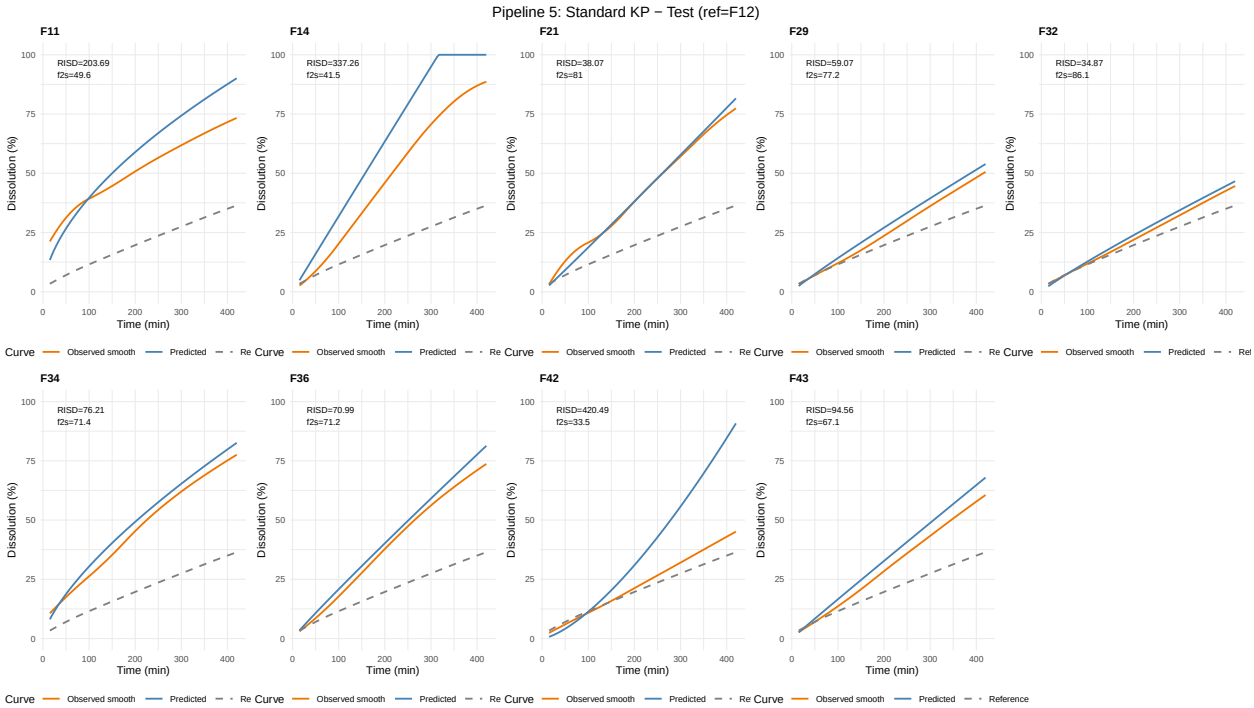


Figure B.12: Full set of reconstructed dissolution curves for Dataset B under the KP model.

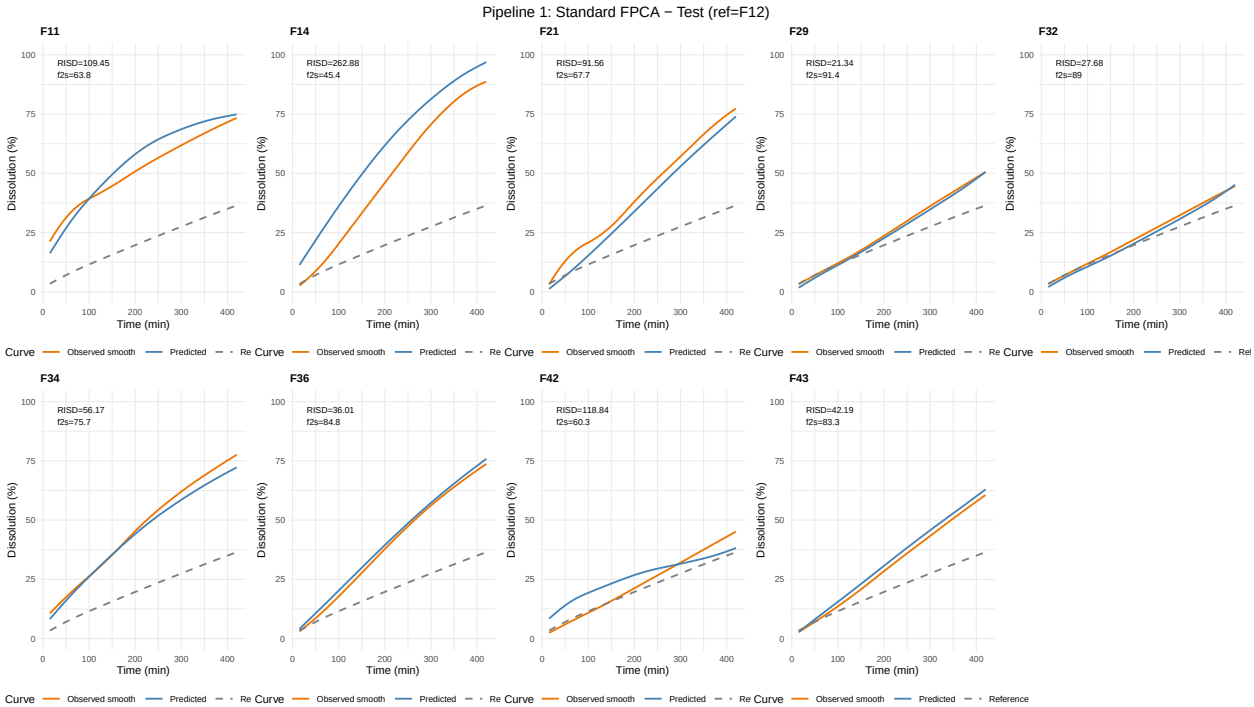


Figure B.13: Full set of reconstructed dissolution curves for Dataset B under the standard FPCA model.

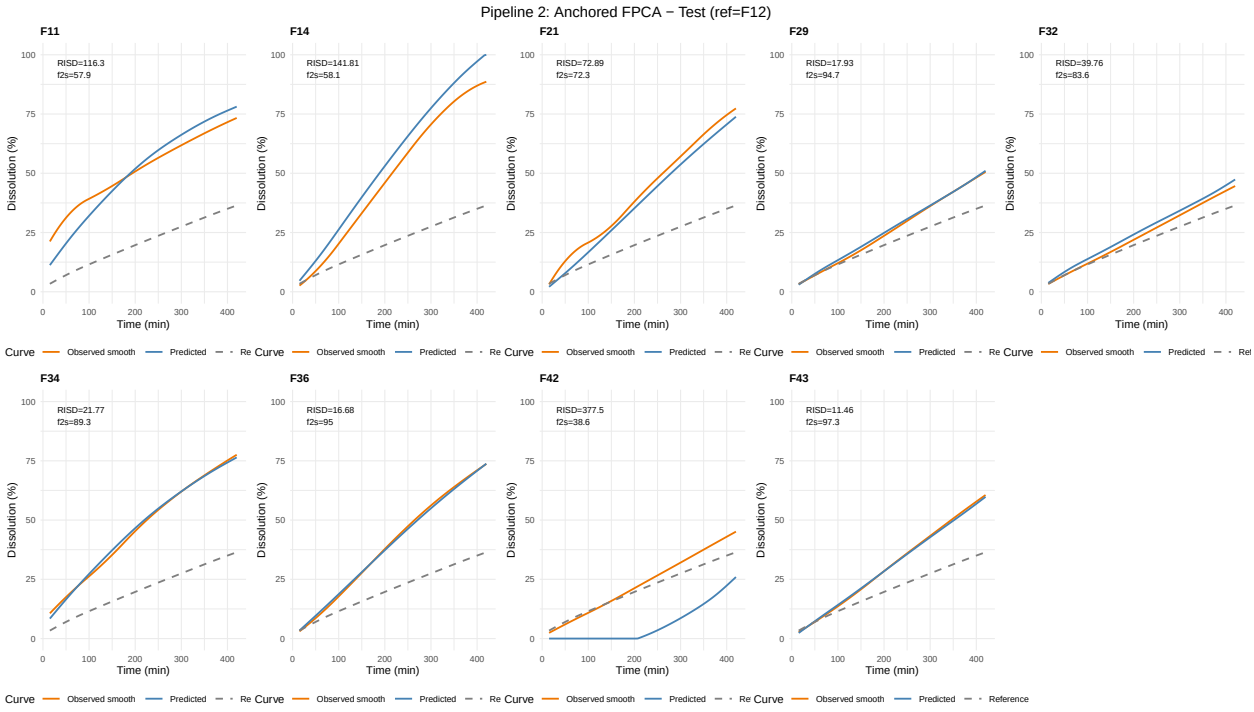


Figure B.14: Full set of reconstructed dissolution curves for Dataset B under the anchored FPCA model.

Variable-selection pipelines

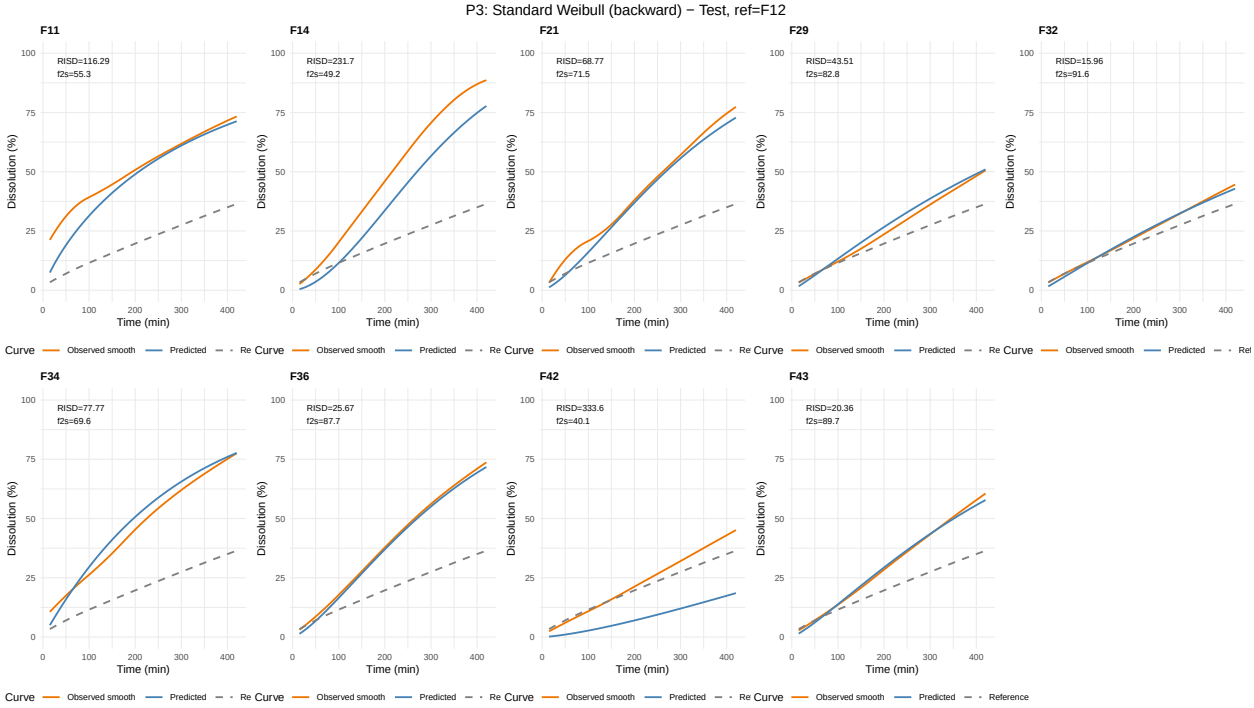


Figure B.15: Reconstructed dissolution curves for Dataset B under the standard Weibull model with variable selection.

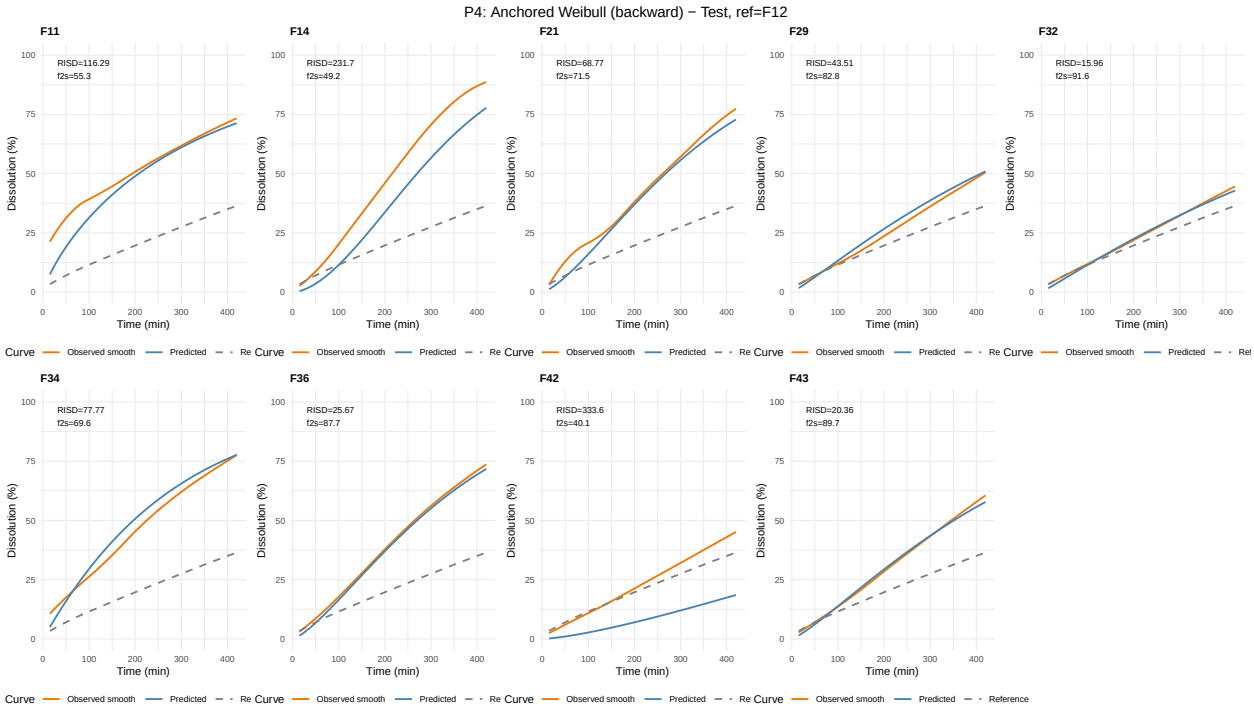


Figure B.16: Reconstructed dissolution curves for Dataset B under the anchored Weibull model with variable selection.

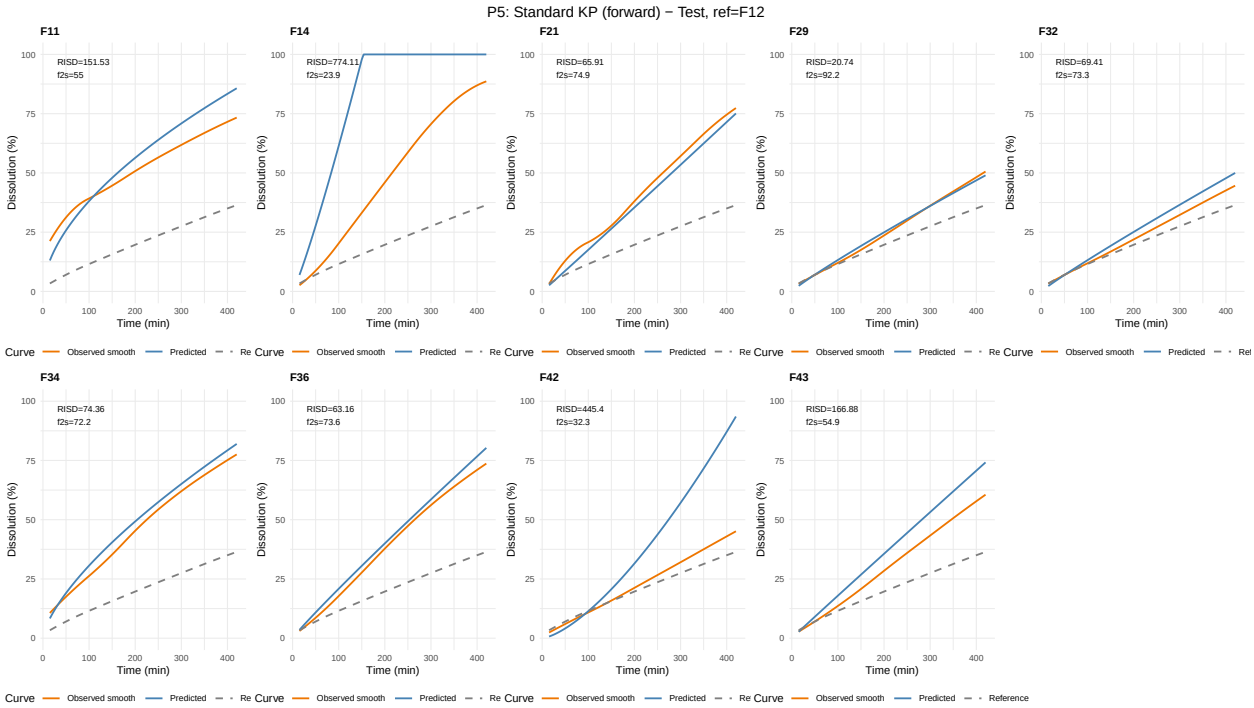


Figure B.17: Reconstructed dissolution curves for Dataset B under the standard KP model with variable selection.

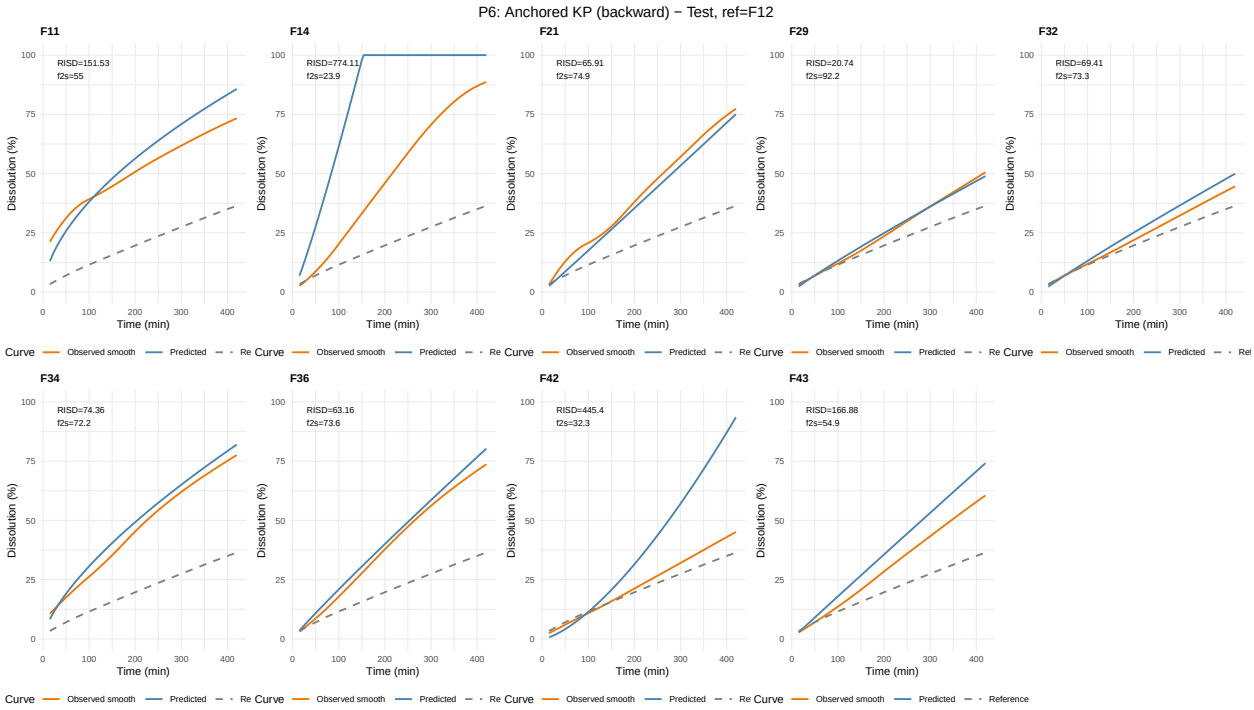


Figure B.18: Reconstructed dissolution curves for Dataset B under the anchored KP model with variable selection.

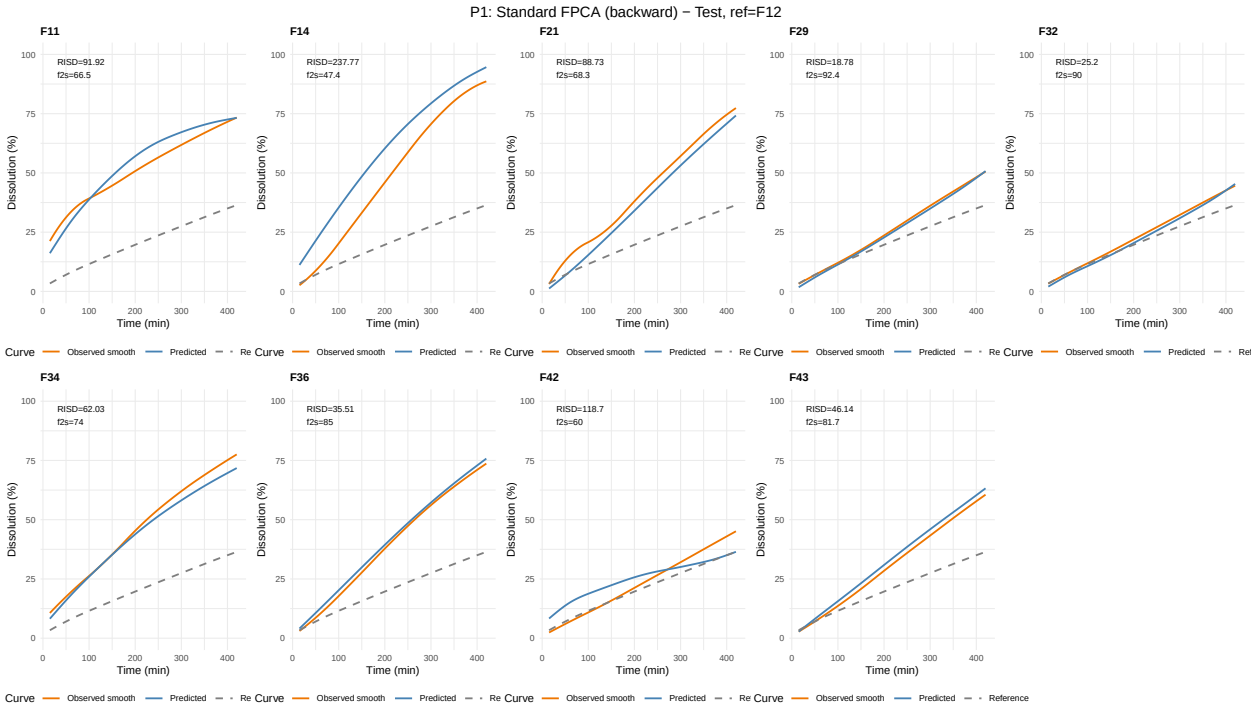


Figure B.19: Reconstructed dissolution curves for Dataset B under the standard FPCA model with variable selection.

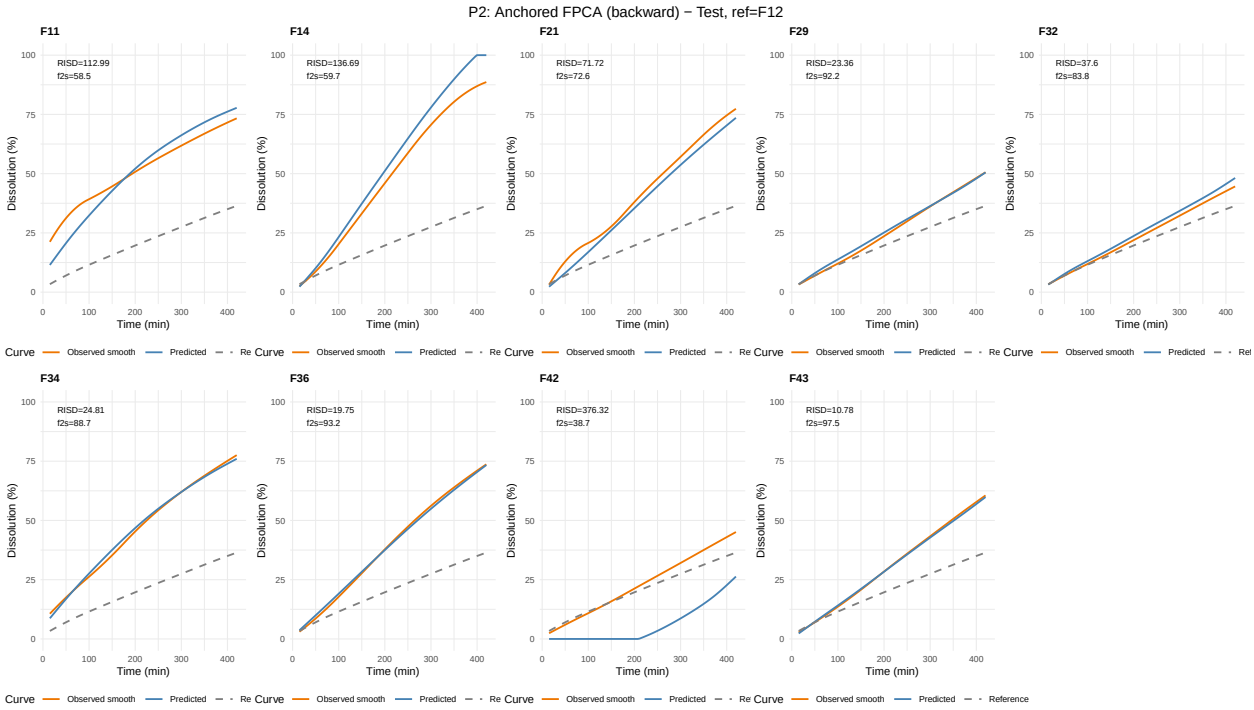


Figure B.20: Reconstructed dissolution curves for Dataset B under the anchored FPCA model with variable selection.

B.2 Optimization Plots

Under the full-model specification, the standard and anchored parametric pipelines may yield identical optimized profiles when the anchoring effect is absorbed by the regression structure. Accordingly, when the resulting top-ranked optimization plots are identical, only one set of plots is presented.

B.2.1 Dataset A

Full-model pipelines

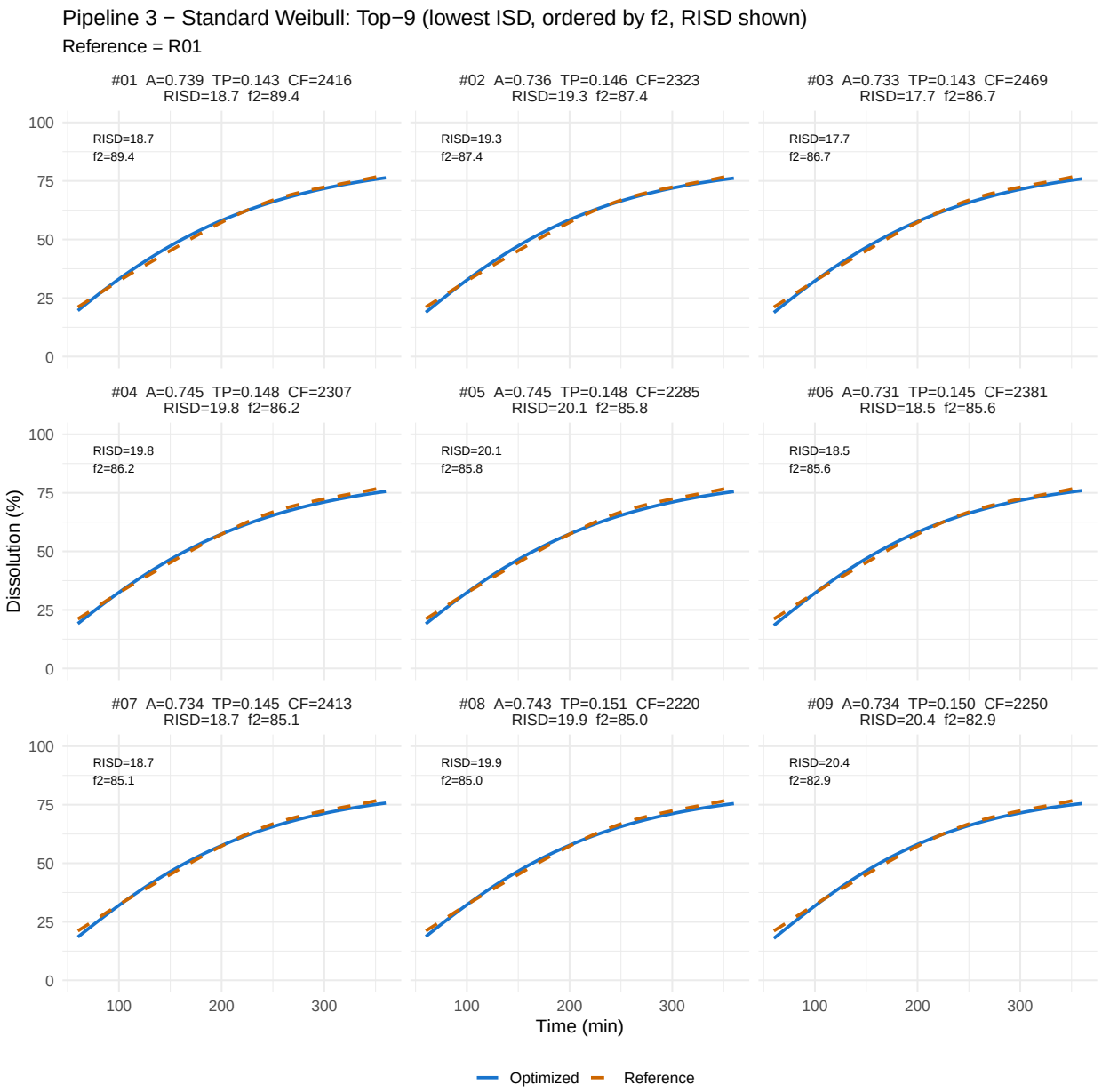


Figure B.21: Top 9 optimized dissolution profiles for Dataset A under the Weibull model. The standard and anchored formulations produce identical results under the full-model specification.

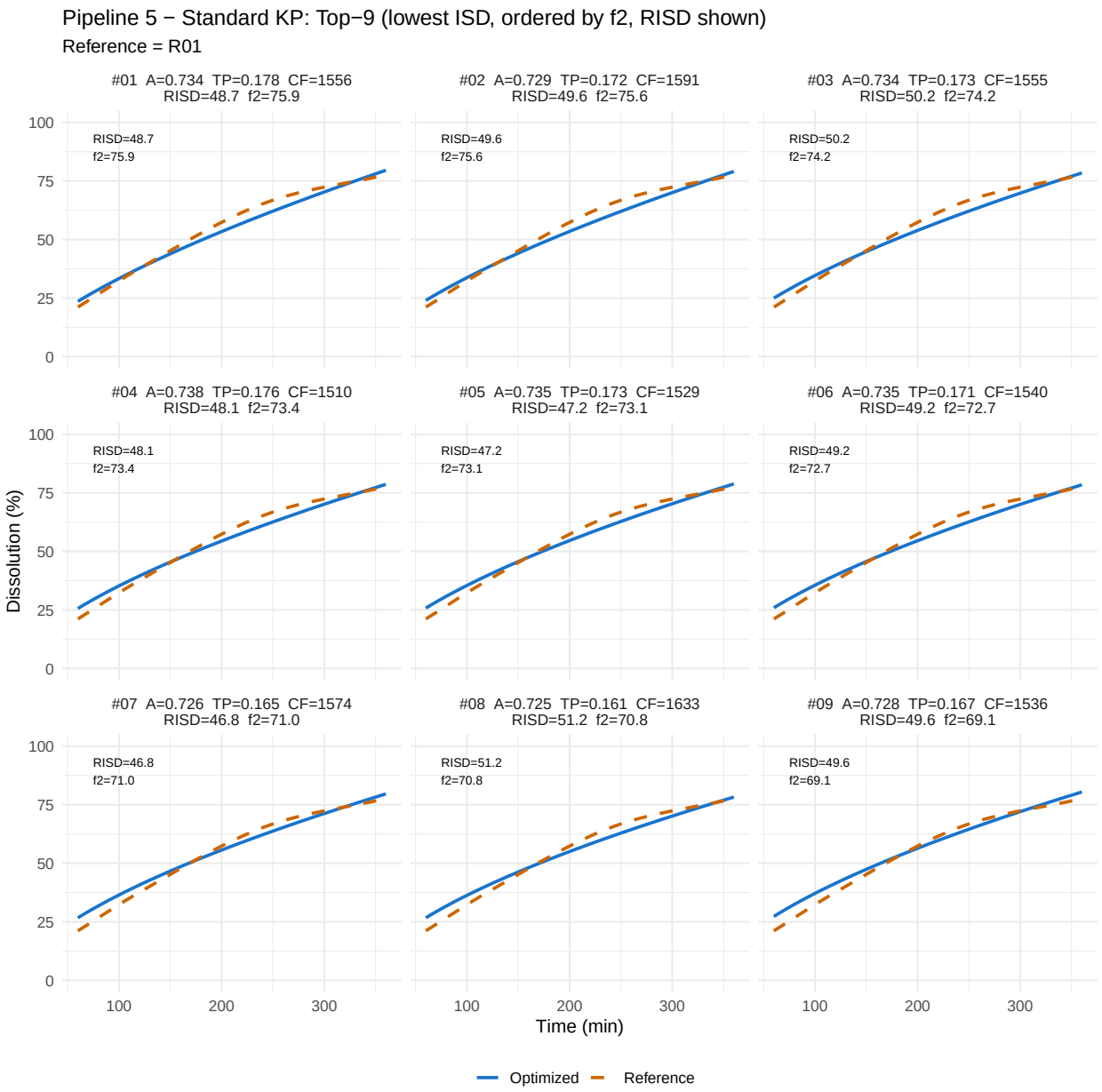


Figure B.22: Top 9 optimized dissolution profiles for Dataset A under the KP model. The standard and anchored formulations produce identical results under the full-model specification.

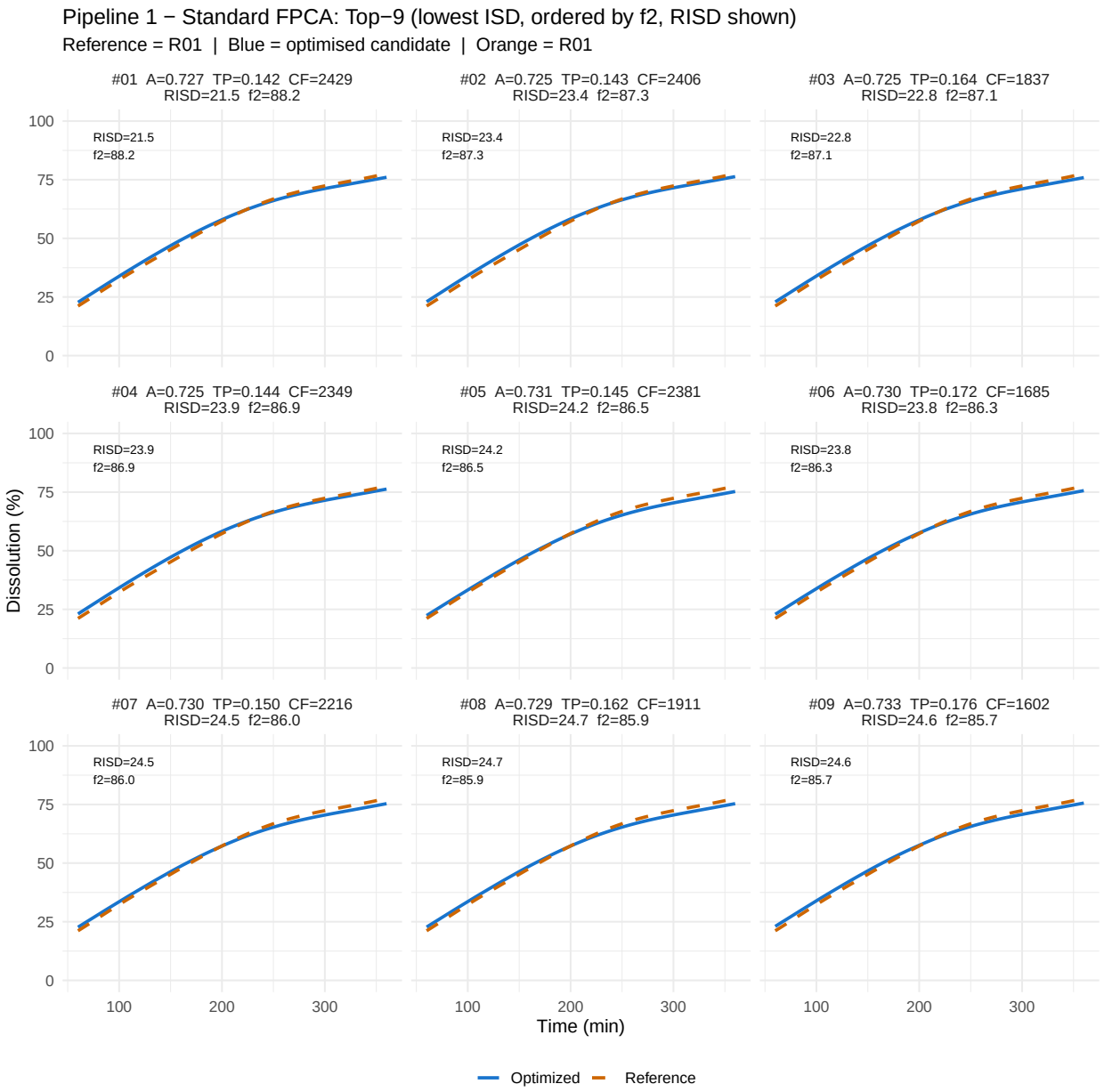


Figure B.23: Top 9 optimized dissolution profiles for Dataset A under the standard FPCA model.

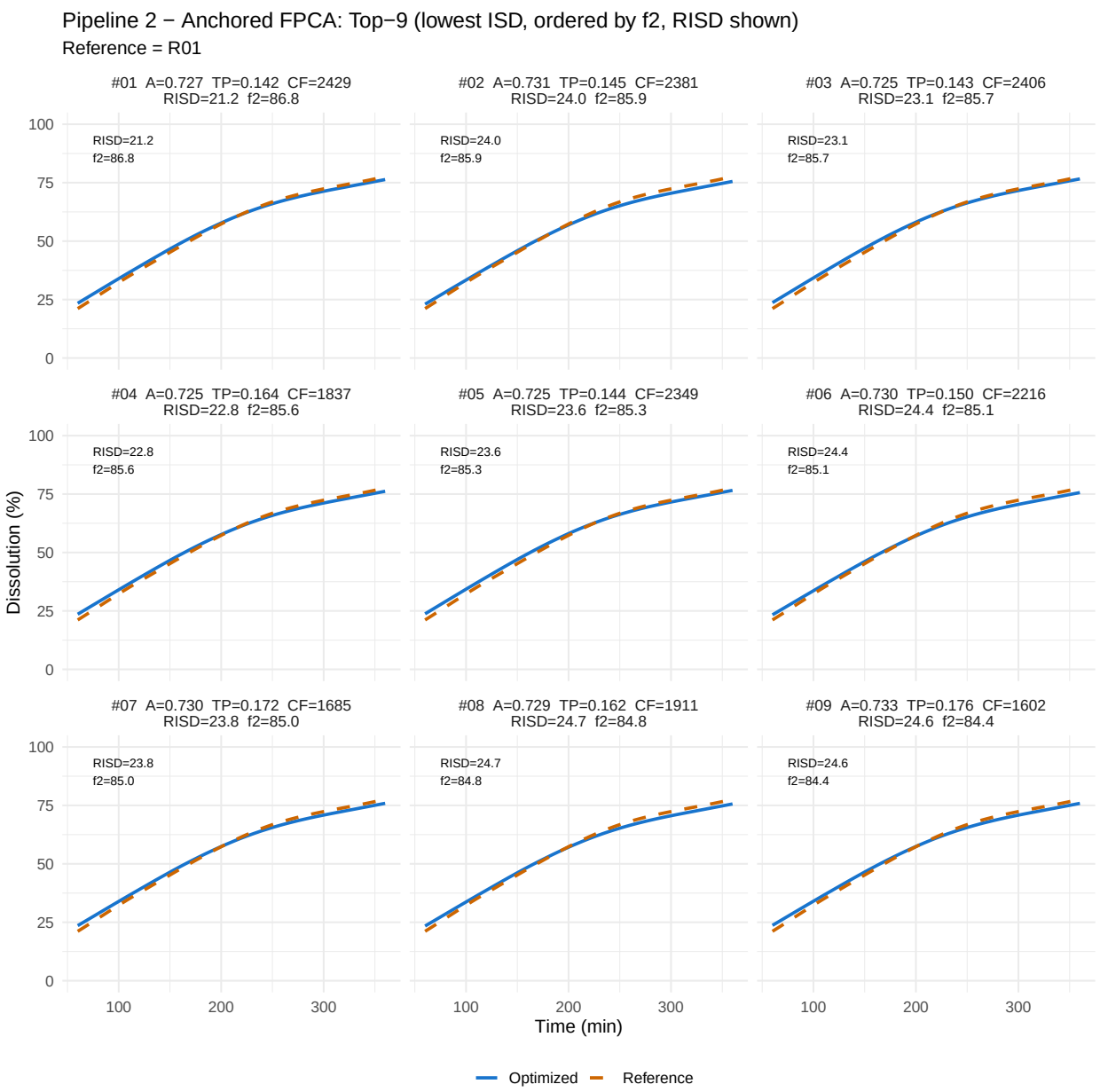


Figure B.24: Top 9 optimized dissolution profiles for Dataset A under the anchored FPCA model.

Variable-selection pipelines

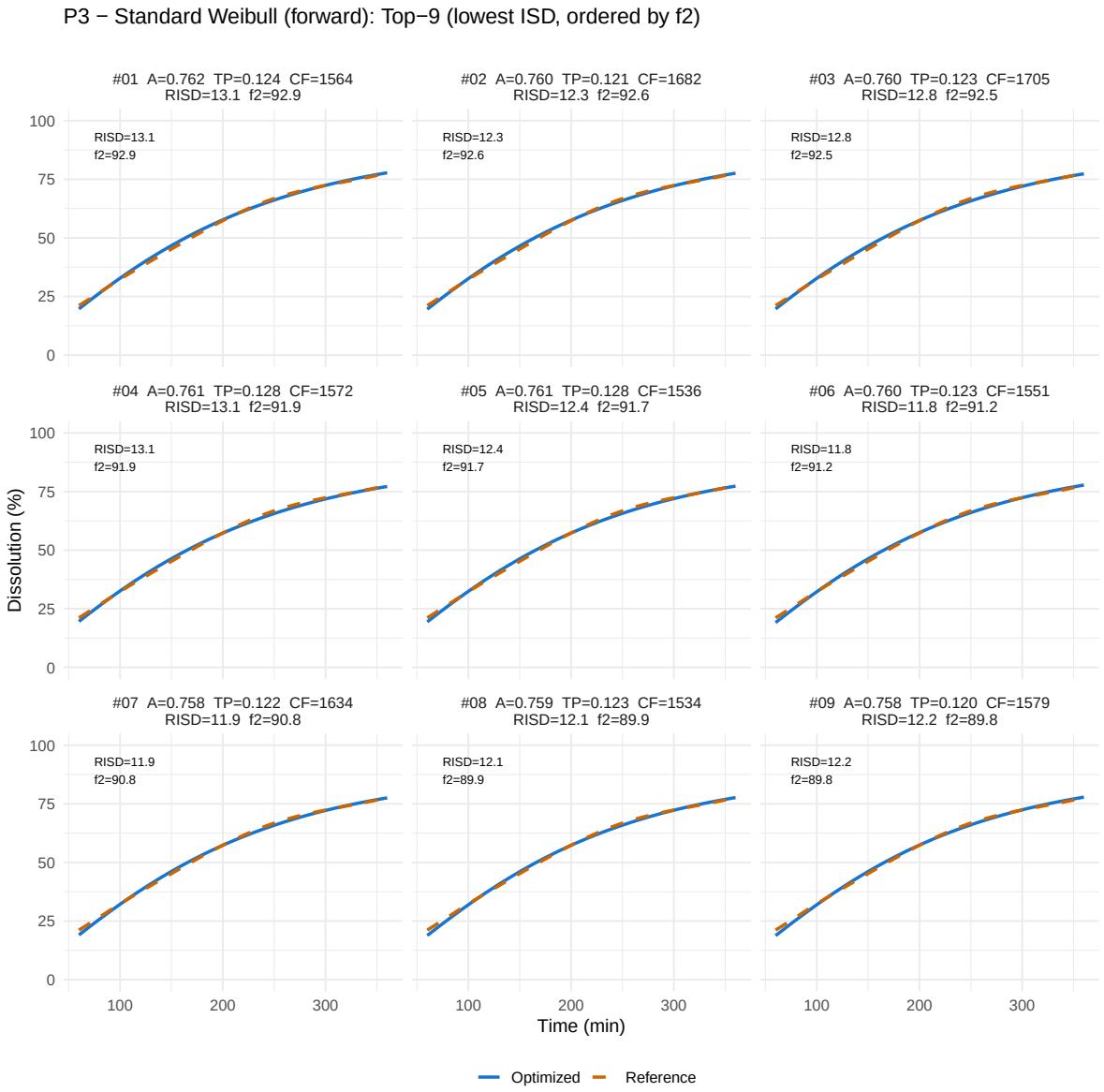


Figure B.25: Top 9 optimized dissolution profiles for Dataset A under the standard Weibull model with variable selection.

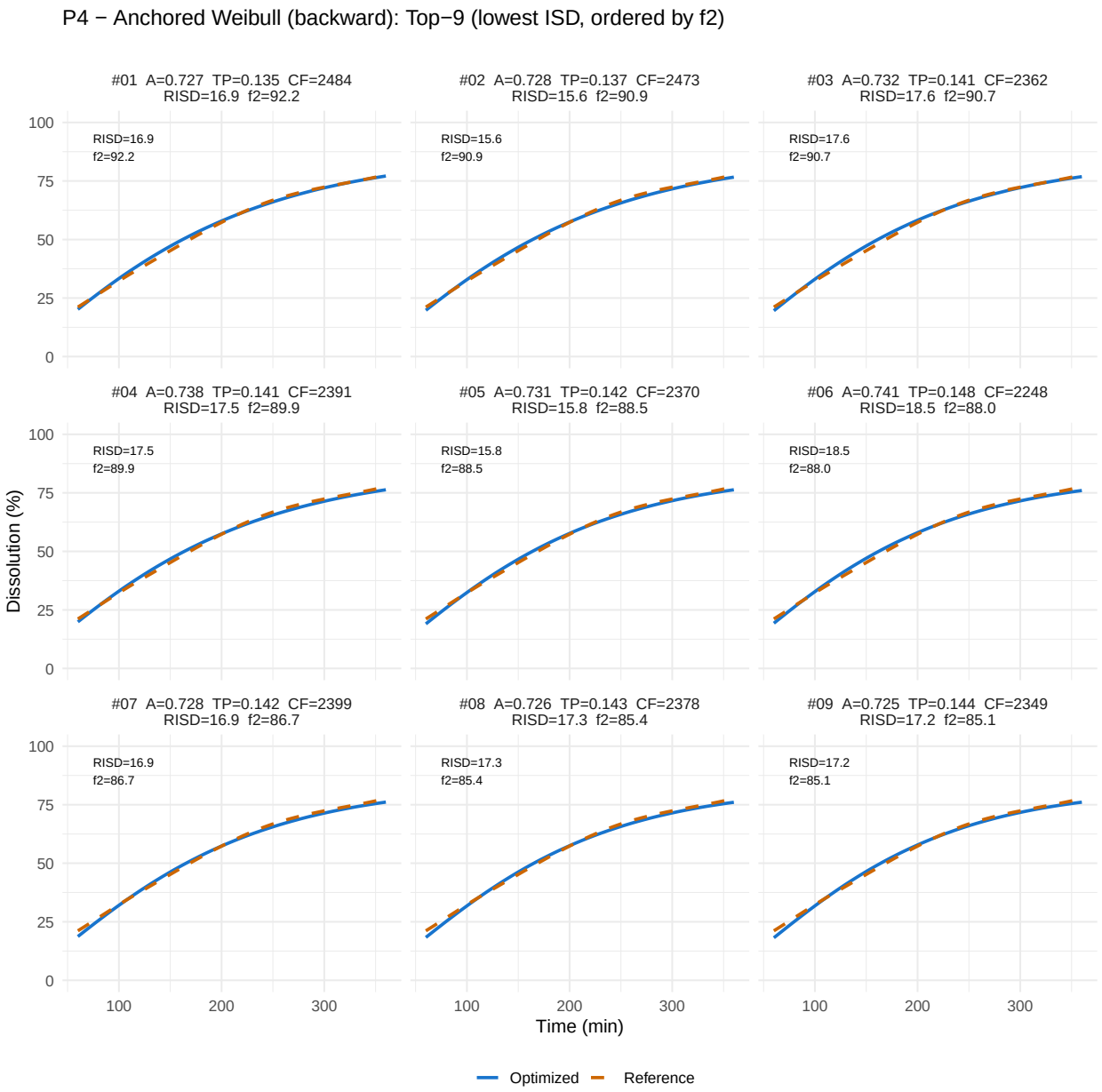


Figure B.26: Top 9 optimized dissolution profiles for Dataset A under the anchored Weibull model with variable selection.

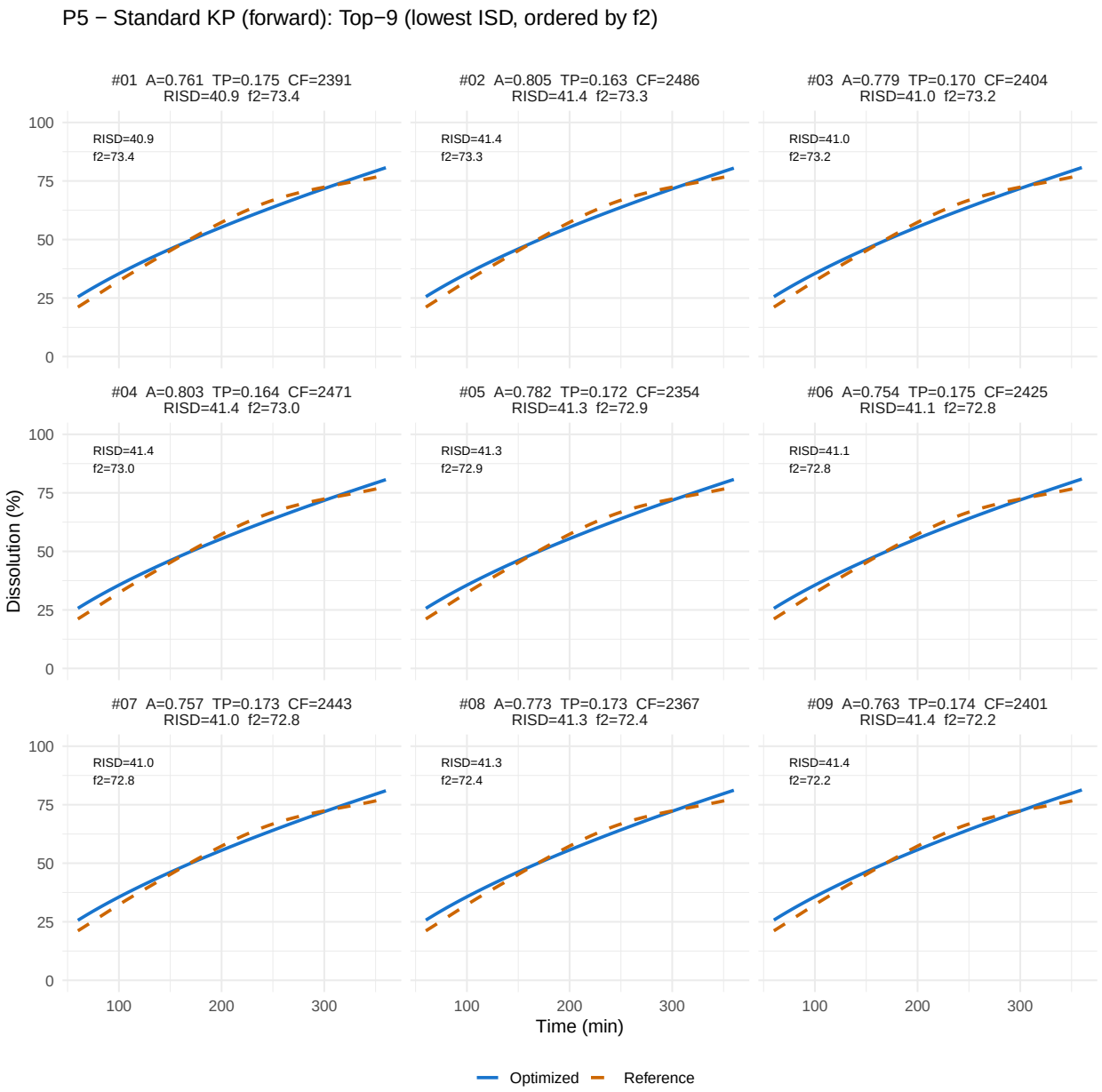


Figure B.27: Top 9 optimized dissolution profiles for Dataset A under the standard KP model with variable selection.

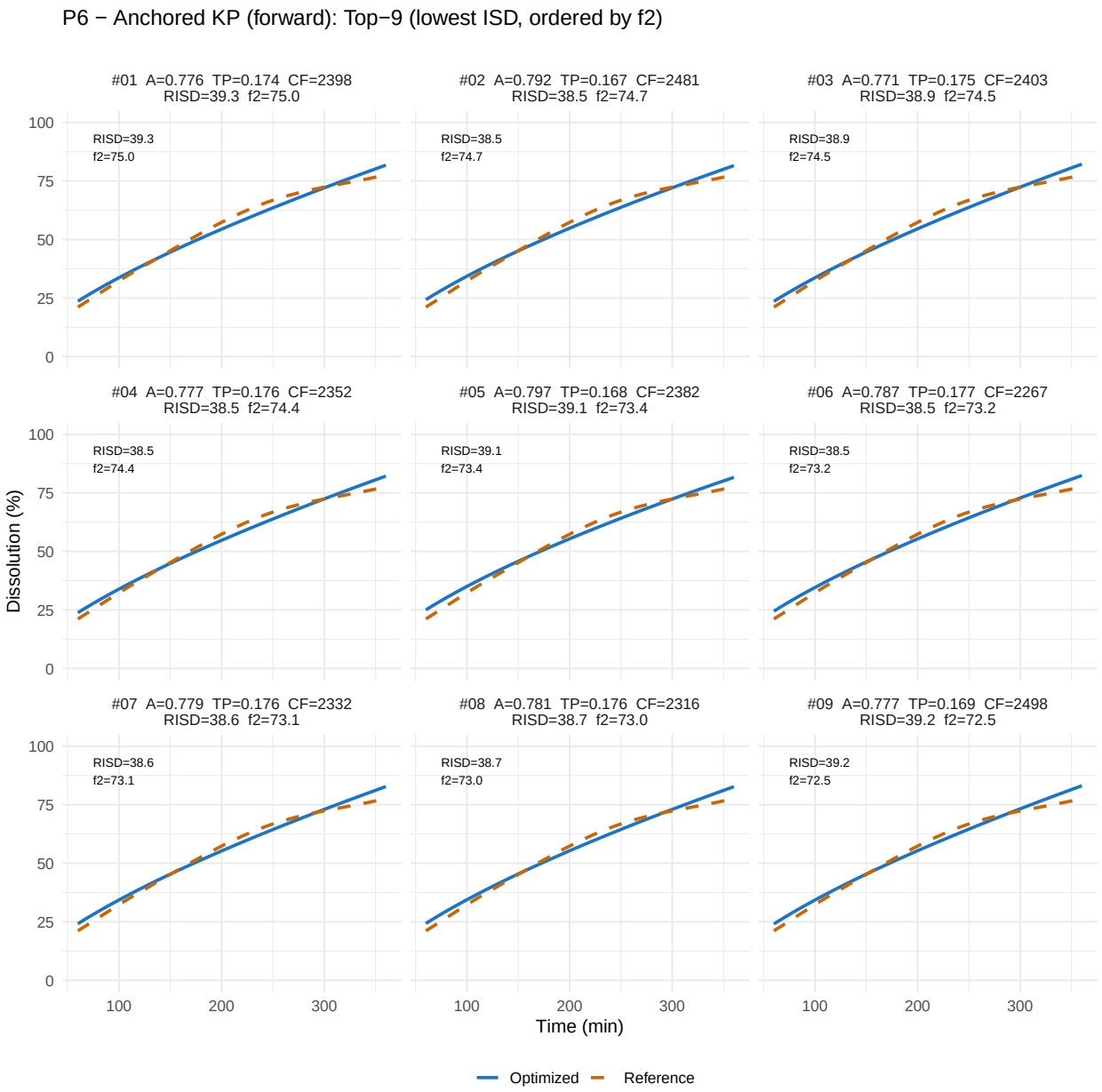


Figure B.28: Top 9 optimized dissolution profiles for Dataset A under the anchored KP model with variable selection.

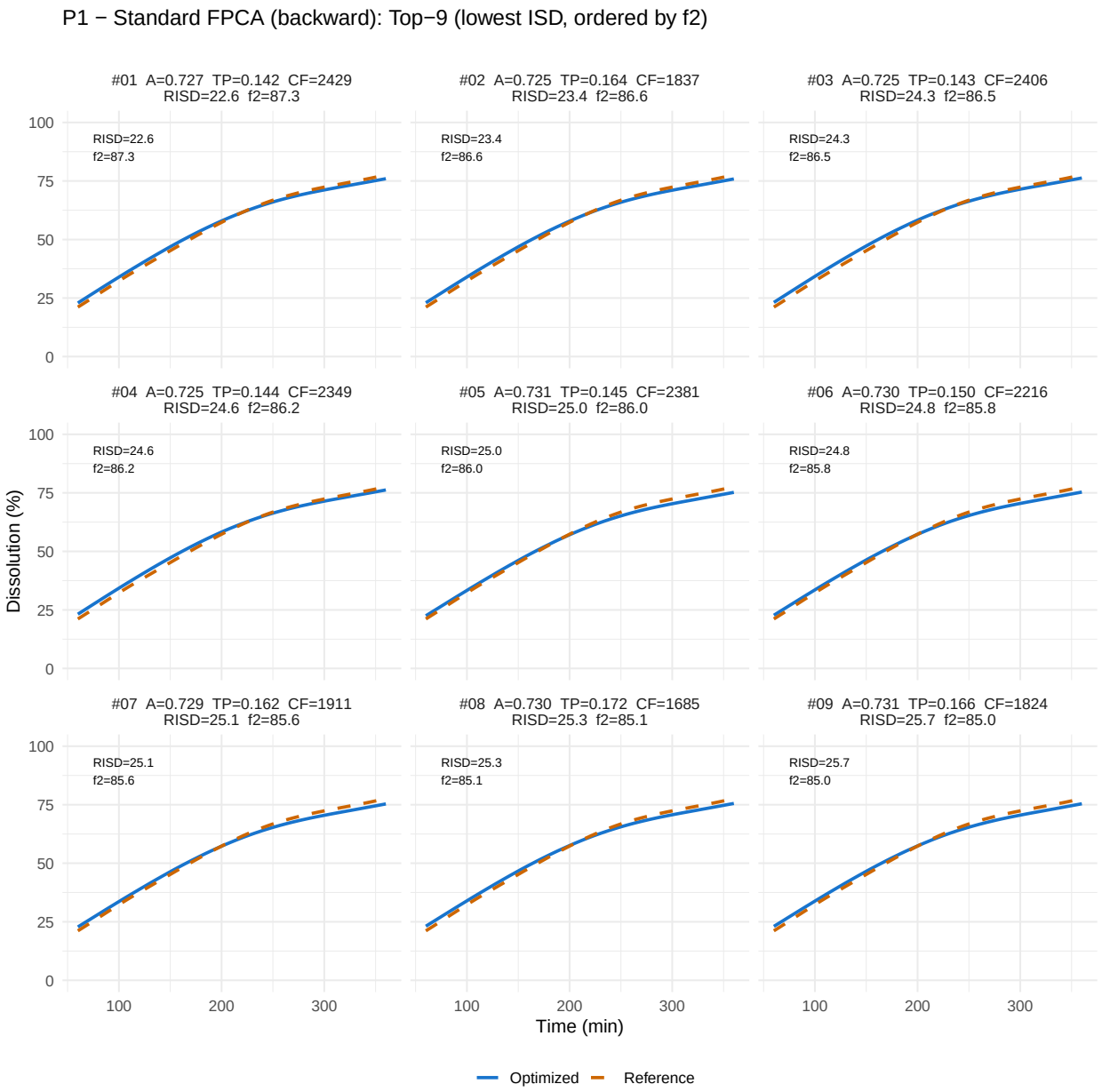


Figure B.29: Top 9 optimized dissolution profiles for Dataset A under the standard FPCA model with variable selection.

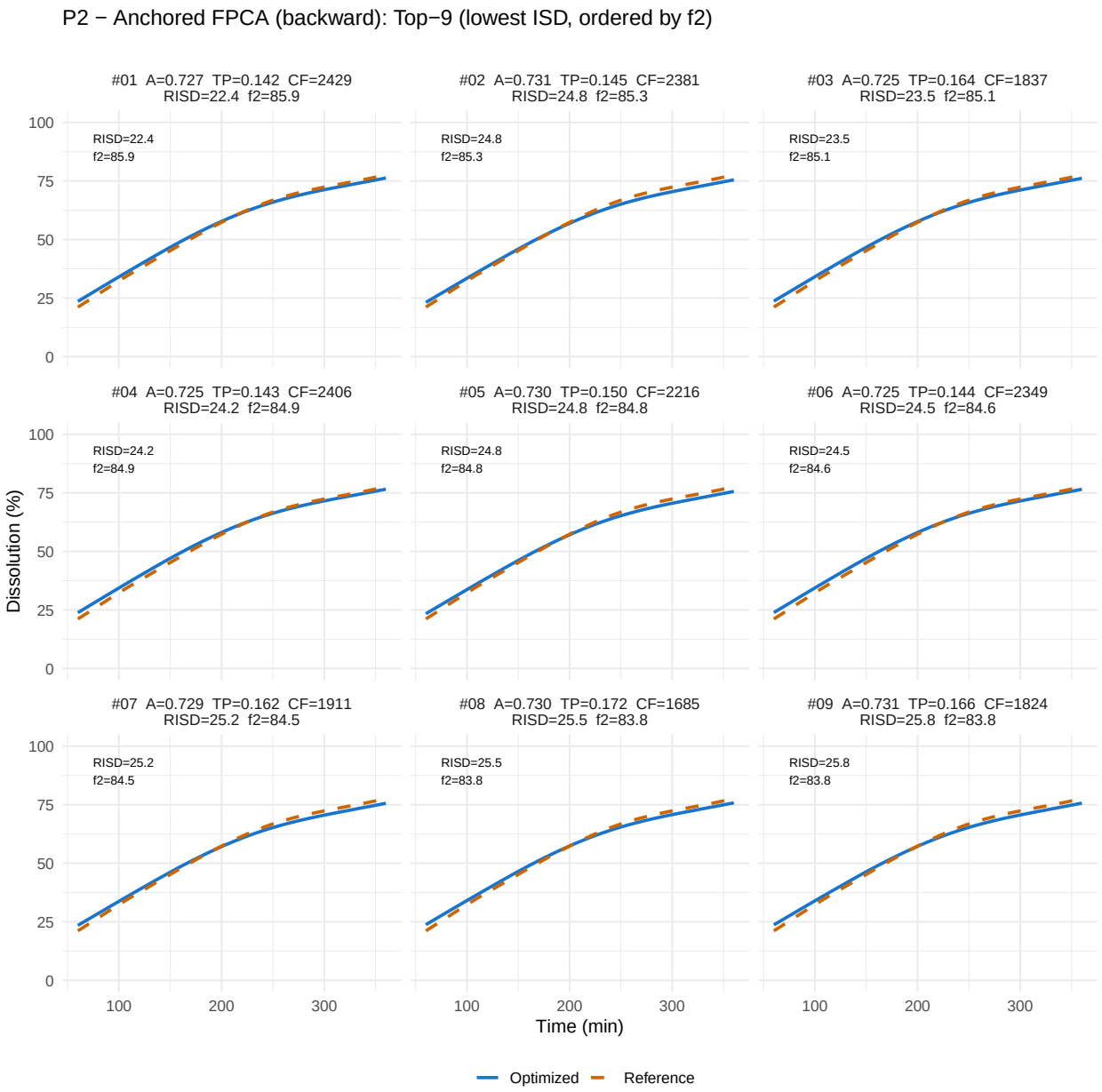


Figure B.30: Top 9 optimized dissolution profiles for Dataset A under the anchored FPCA model with variable selection.

B.2.2 Dataset B

Full-model pipelines

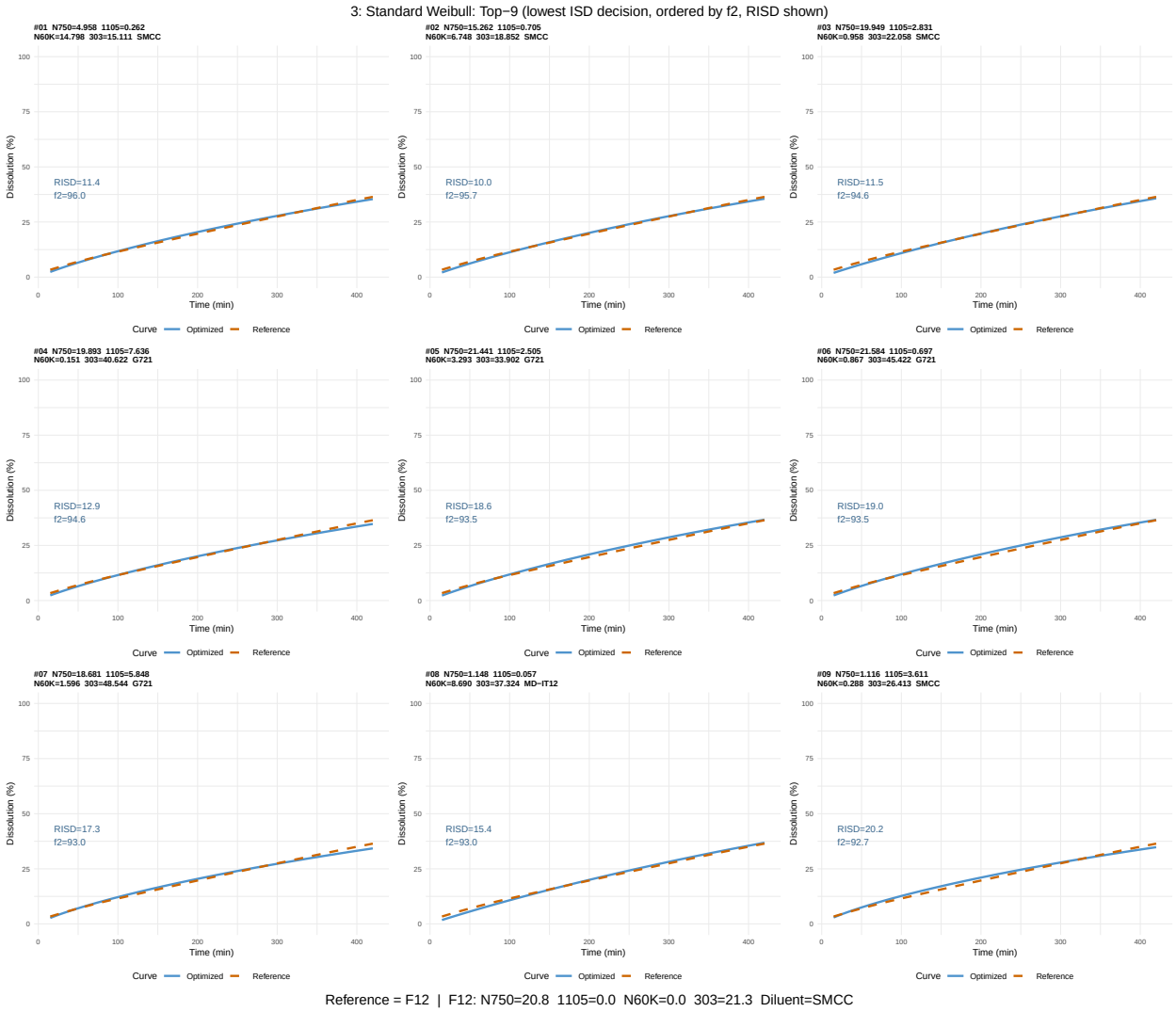


Figure B.31: Top 9 optimized dissolution profiles for Dataset B under the Weibull model. When the standard and anchored formulations are identical under the full-model specification, only one set of plots is shown.

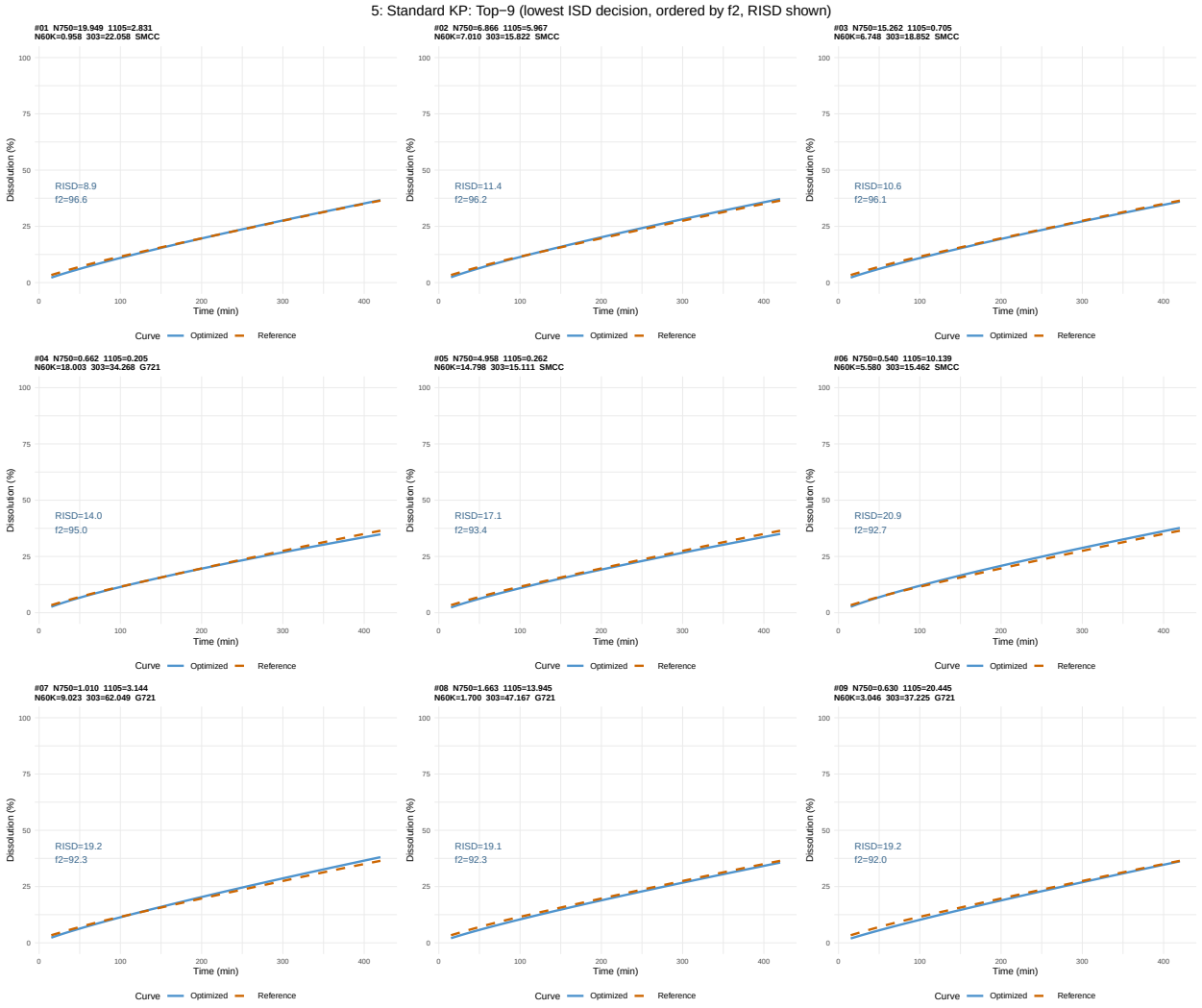


Figure B.32: Top 9 optimized dissolution profiles for Dataset B under the KP model. When the standard and anchored formulations are identical under the full-model specification, only one set of plots is shown.

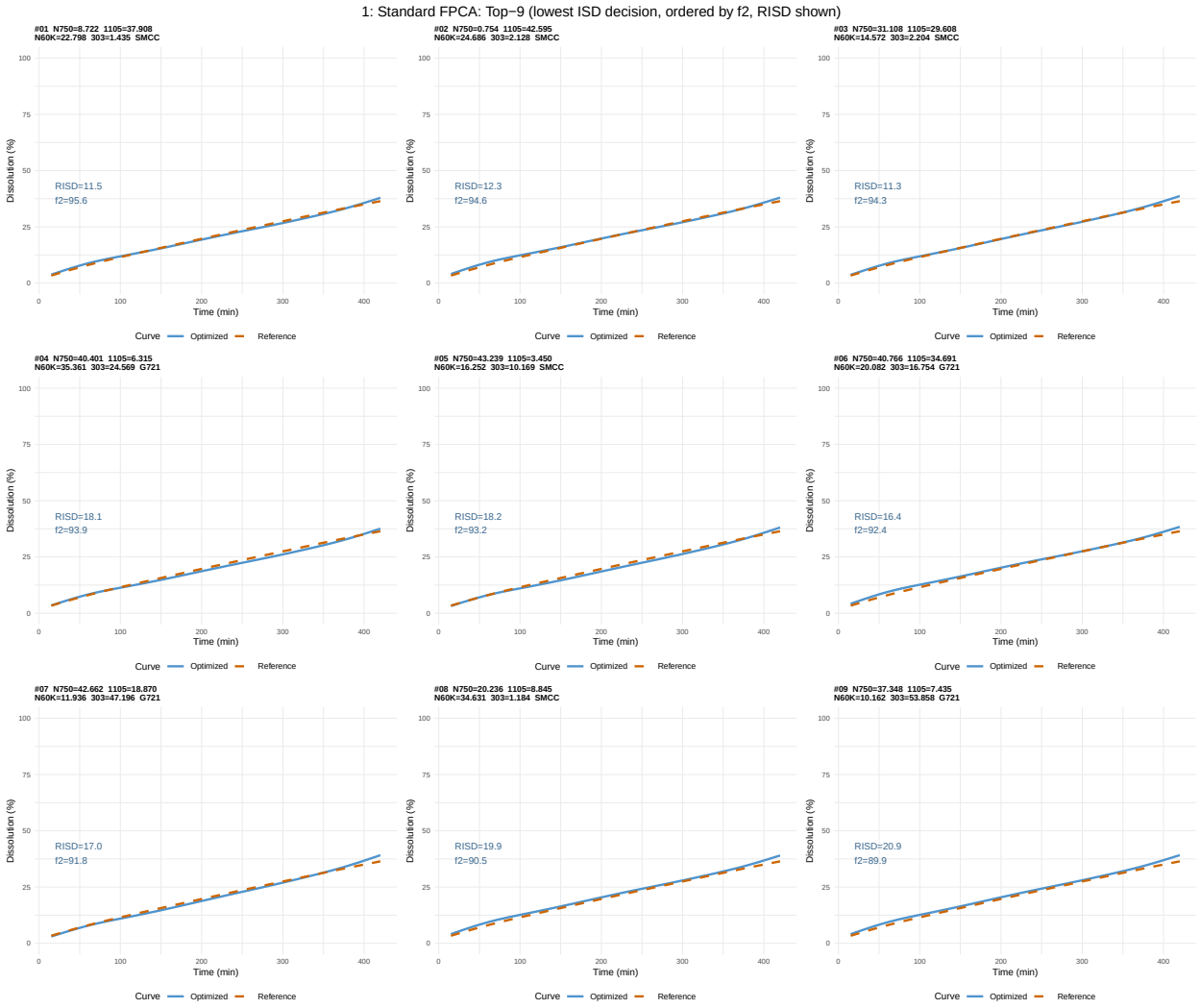


Figure B.33: Top 9 optimized dissolution profiles for Dataset B under the standard FPCA model.

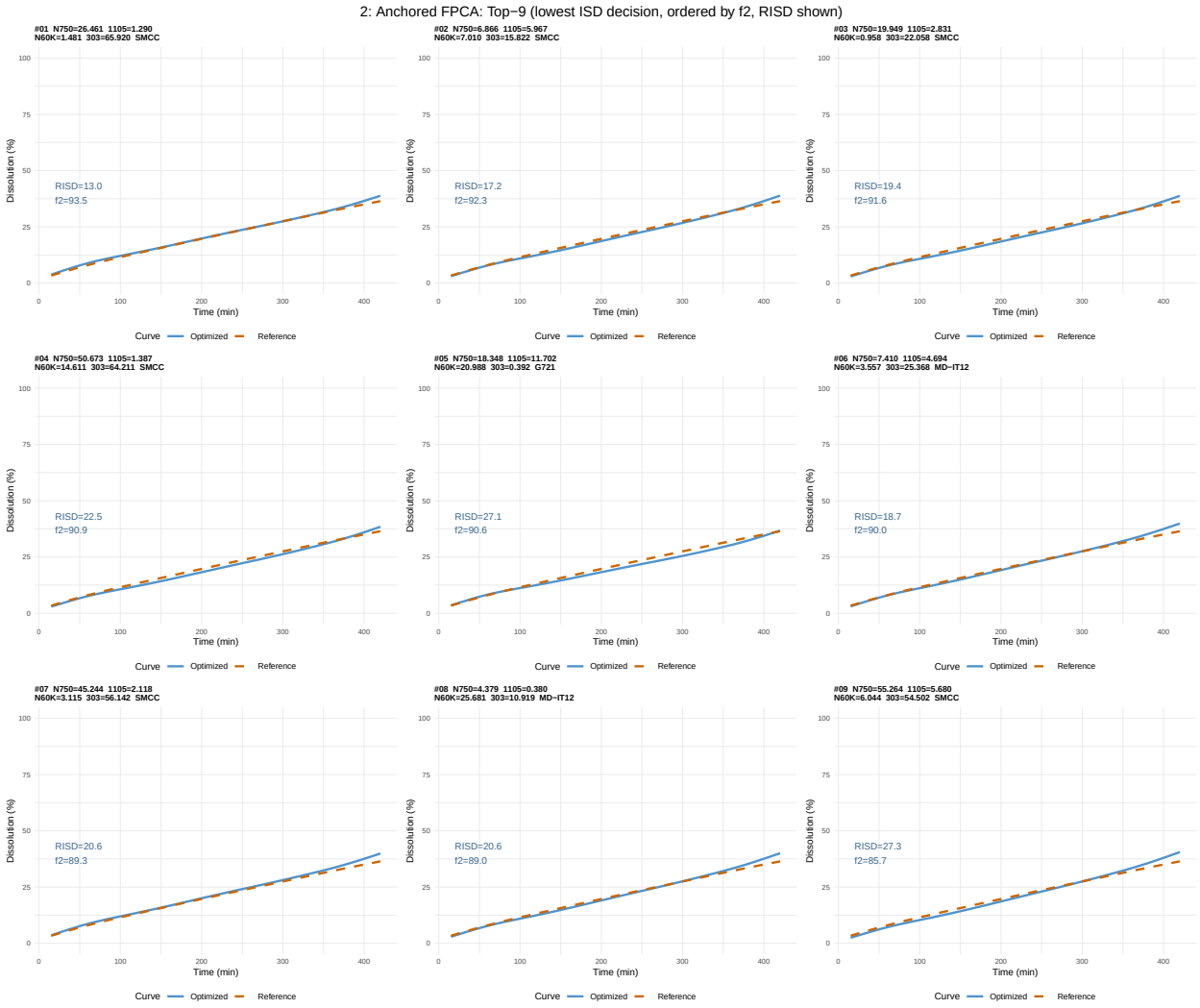


Figure B.34: Top 9 optimized dissolution profiles for Dataset B under the anchored FPCA model.

Variable-selection pipelines

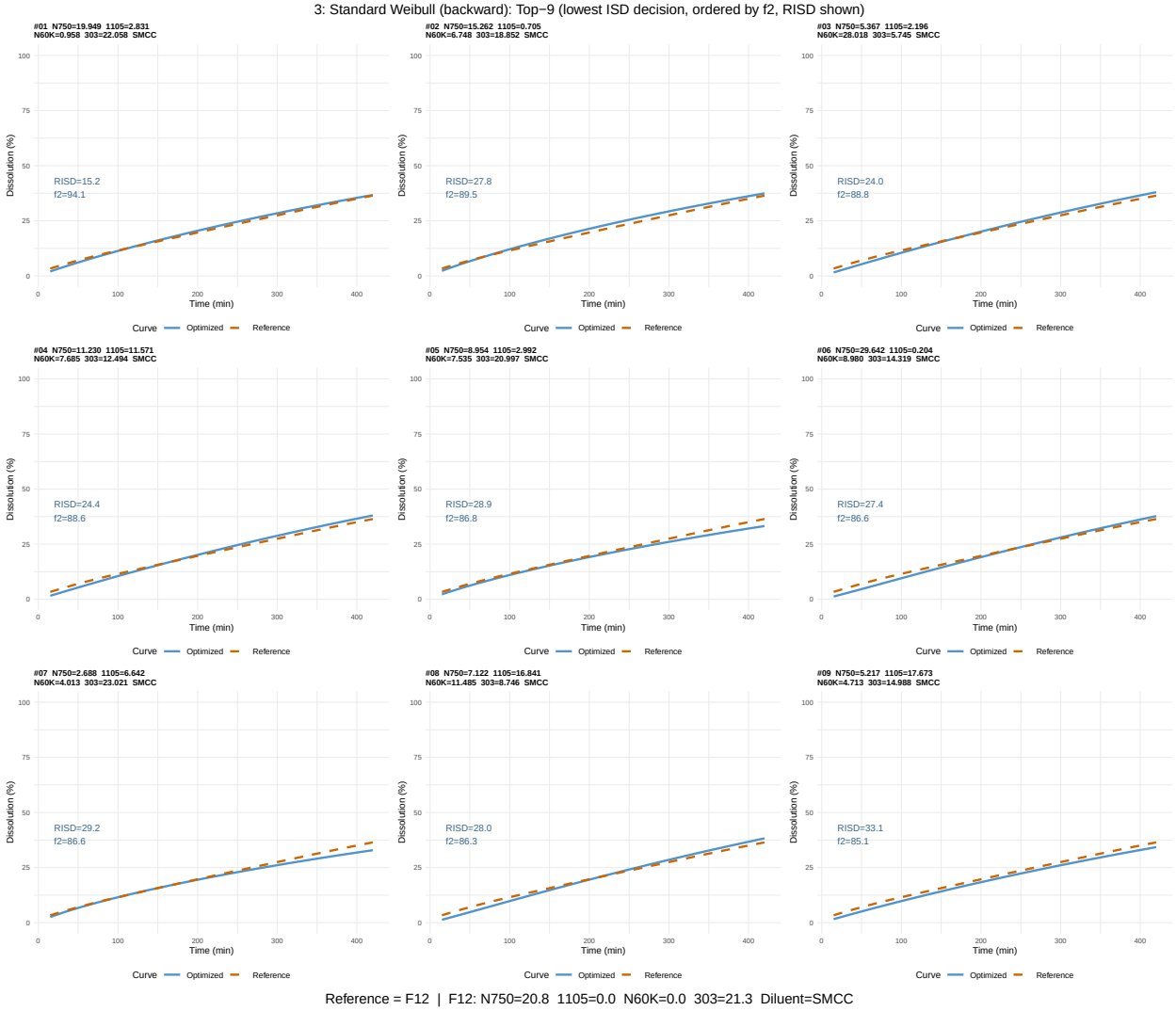


Figure B.35: Top 9 optimized dissolution profiles for Dataset B under the standard Weibull model with variable selection.

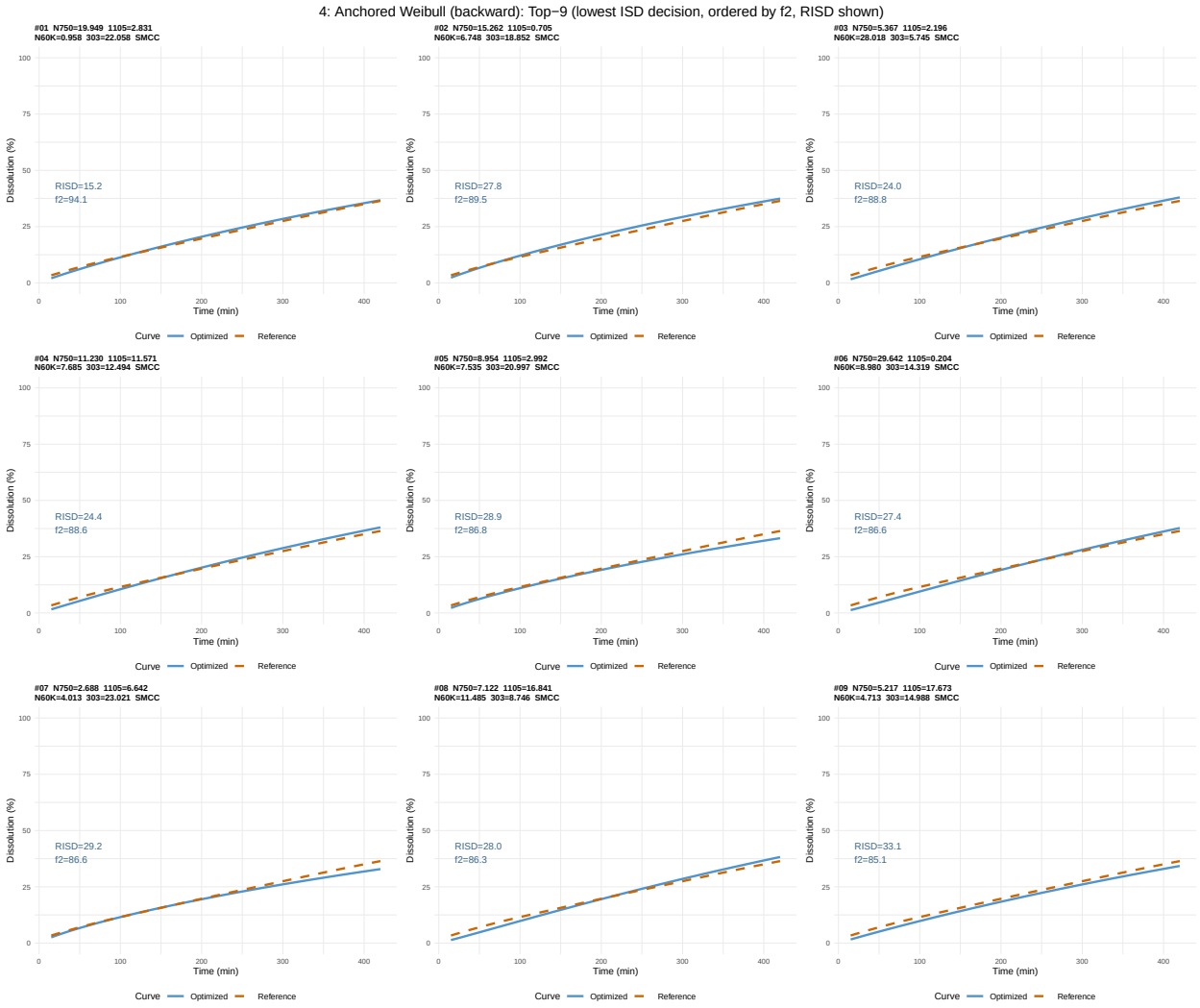


Figure B.36: Top 9 optimized dissolution profiles for Dataset B under the anchored Weibull model with variable selection.

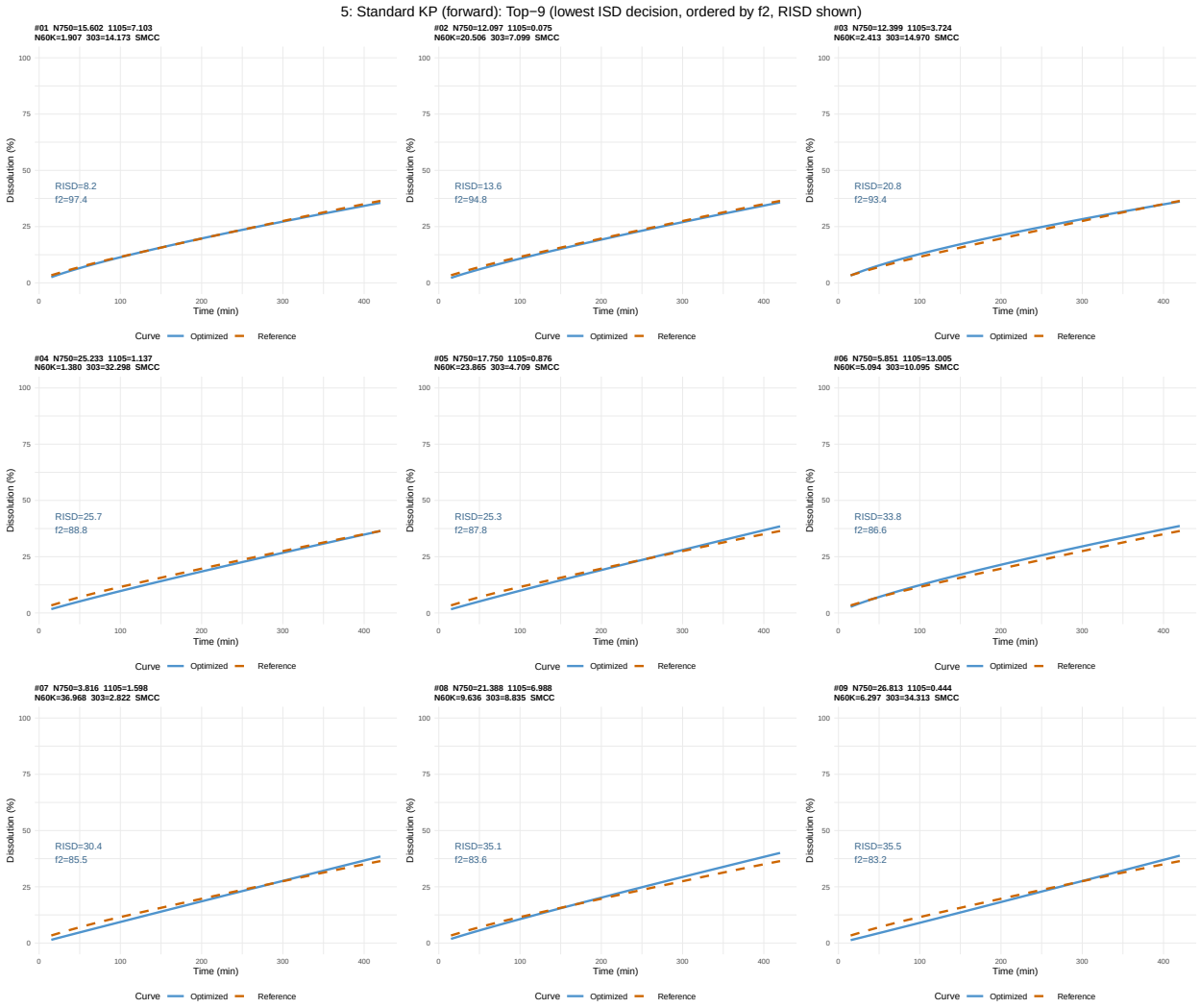


Figure B.37: Top 9 optimized dissolution profiles for Dataset B under the standard KP model with variable selection.

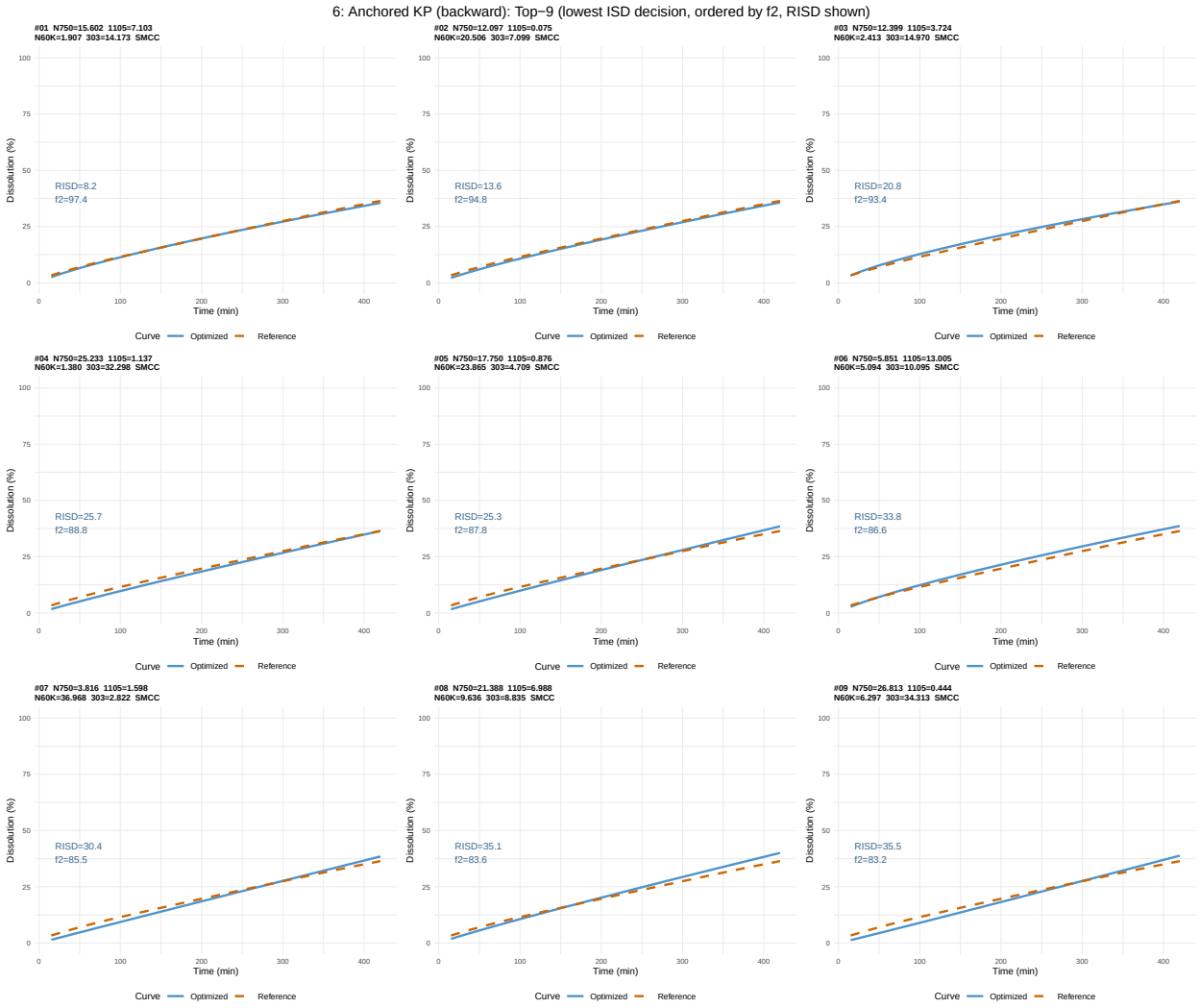


Figure B.38: Top 9 optimized dissolution profiles for Dataset B under the anchored KP model with variable selection.

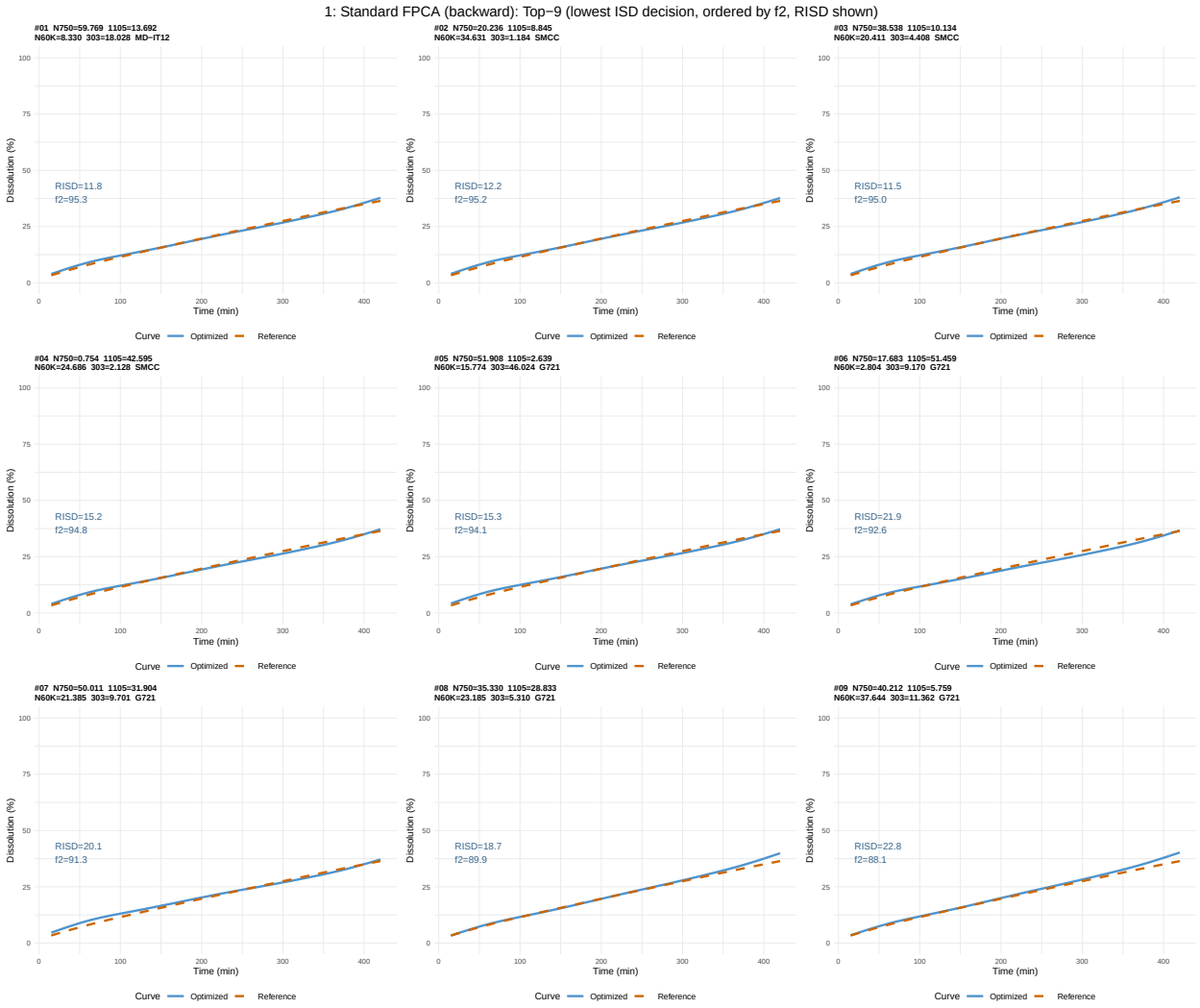


Figure B.39: Top 9 optimized dissolution profiles for Dataset B under the standard FPCA model with variable selection.

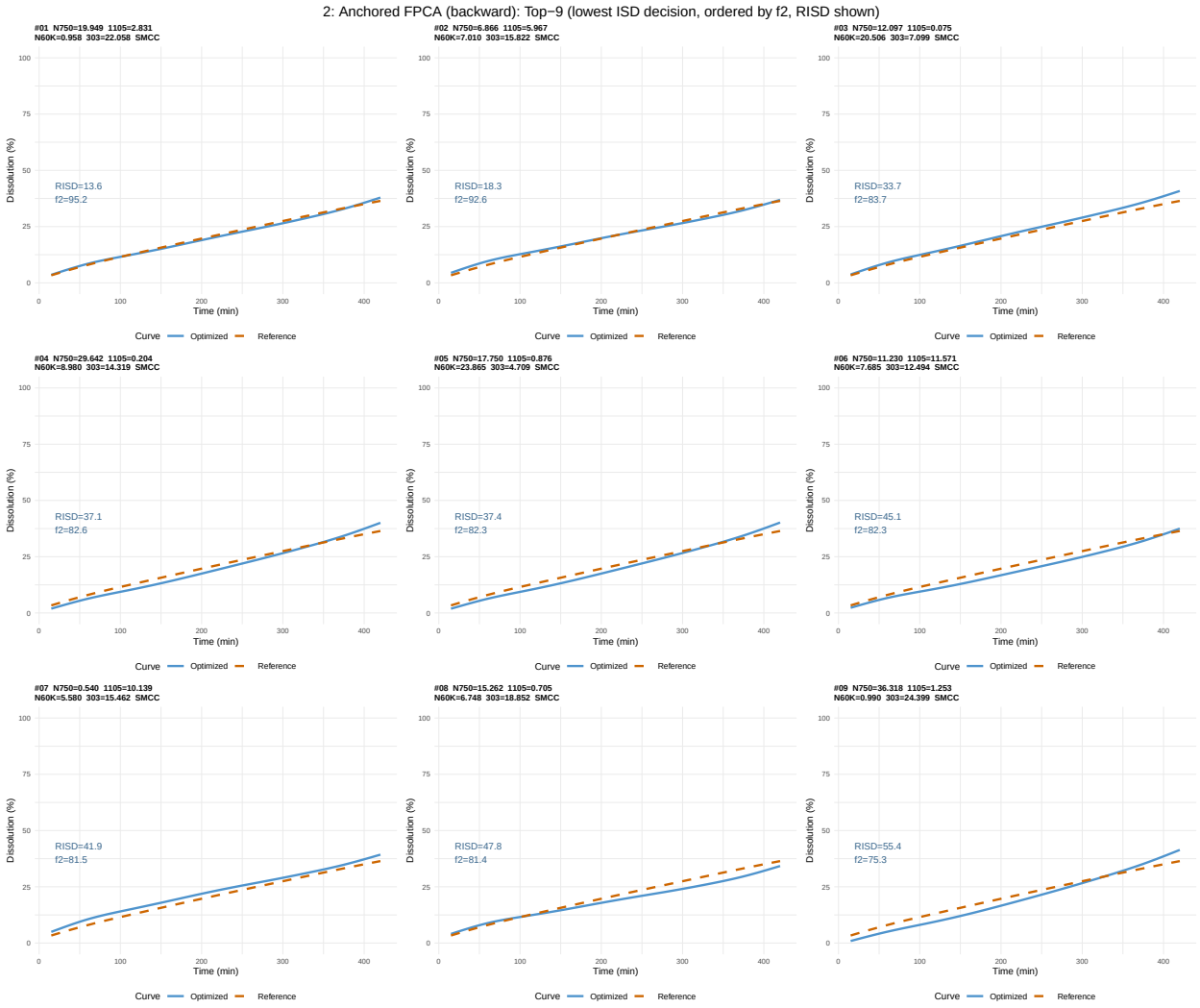
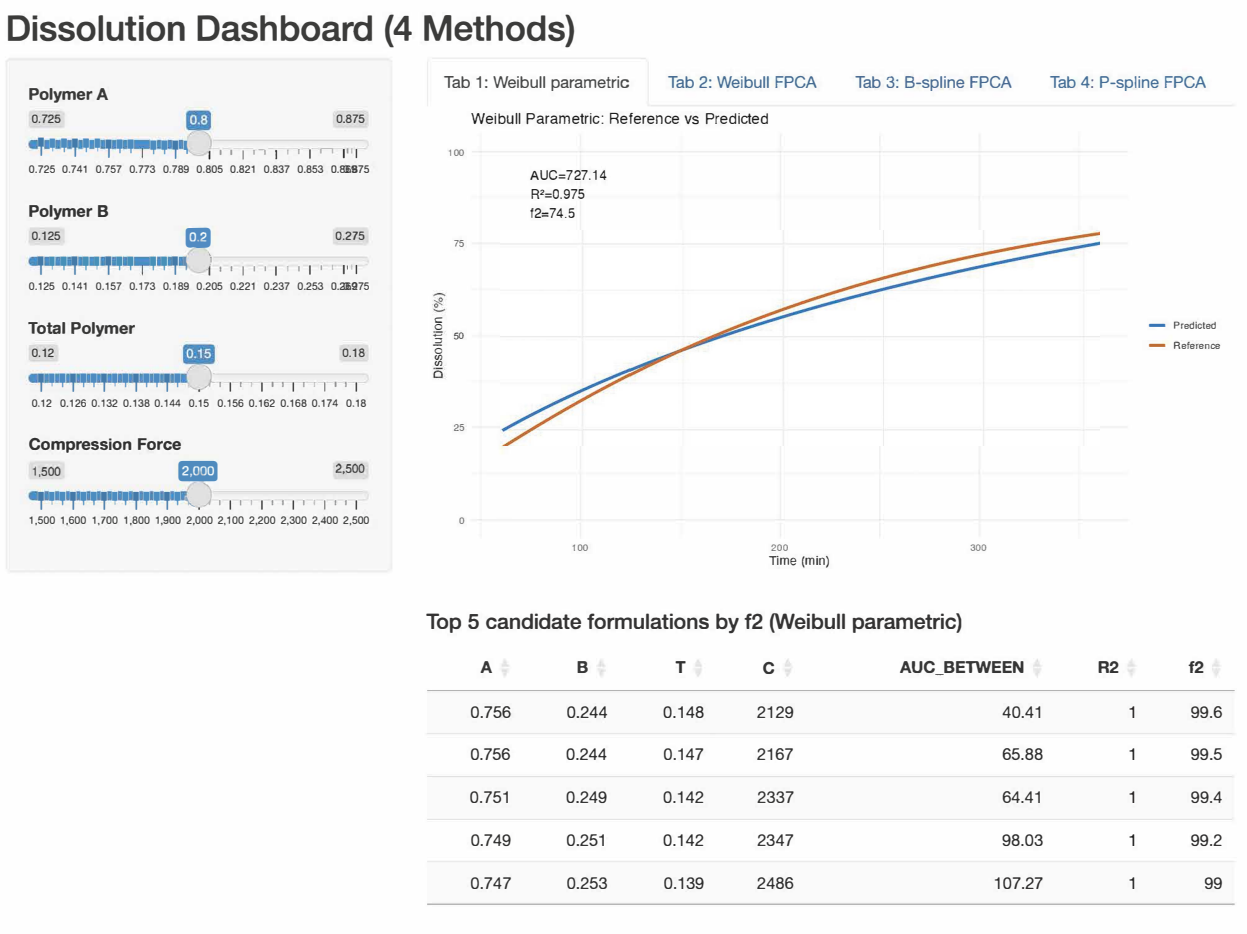


Figure B.40: Top 9 optimized dissolution profiles for Dataset B under the anchored FPCA model with variable selection.

B.3 Dashboard Interface

To demonstrate the practical implementation of the proposed methodology, a dashboard interface was developed for interactive visualization and exploration of dissolution profiles and optimization results.

Figure B.41 provides a representative screenshot of the dashboard, illustrating key components including the reference profile, optimized candidate curves, and associated similarity metrics.



A	B	T	C	AUC_BETWEEN	R2	f2
0.756	0.244	0.148	2129	40.41	1	99.6
0.756	0.244	0.147	2167	65.88	1	99.5
0.751	0.249	0.142	2337	64.41	1	99.4
0.749	0.251	0.142	2347	98.03	1	99.2
0.747	0.253	0.139	2486	107.27	1	99

Figure B.41: Representative screenshot of the interactive dashboard for dissolution profile analysis and formulation optimization. The interface displays the reference curve, optimized candidates, and corresponding similarity metrics (e.g., R^2 and f_2).