Research Paper

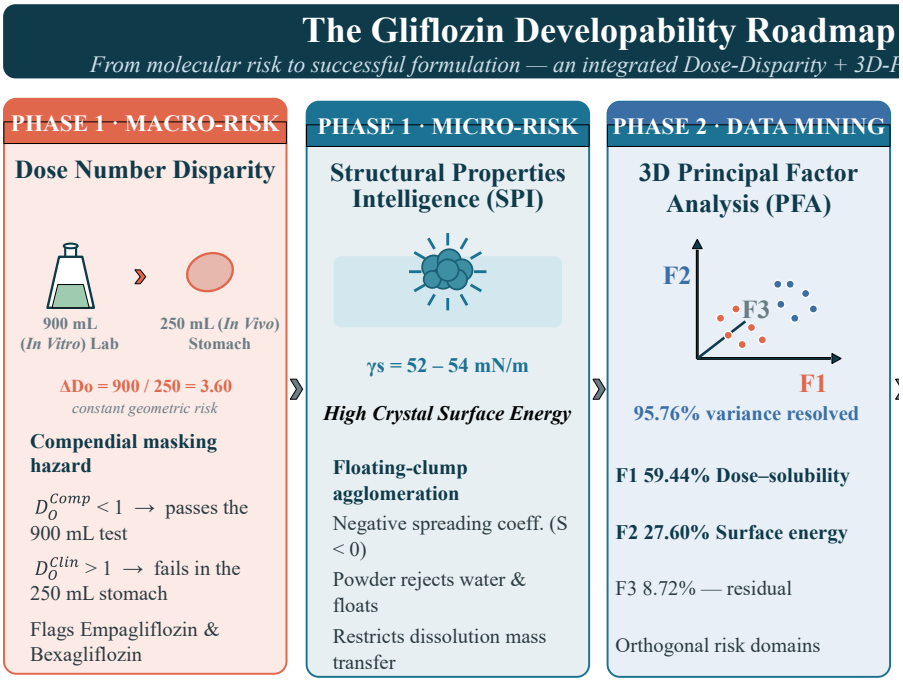
Predictive Developability of Gliflozins: Integrating Dose Number Disparity and Structural Properties Intelligence by Principal Factor Analysis

Anbalagan Uthayathevan[1,*], Dr Ashish Sarkar[1], Chozharajan Tharmaraj[1], Jayaprakash N S[1]
[1]Department of Pharmacy, YBN University, Rajaulatu, Namkum Ranchi, Jharkhand 834010, India
[*]**Corresponding Authors**
1. Anbalagan Uthayathevan, uthayanbupharm1@gmail.com

ABSTRACT
Development is often limited not by potency but by biopharmaceutical and solid-state behaviour. Conventional developability screens rarely capture how dose, gastrointestinal fluid volume and crystal surface energetics interact. We describe a framework combining Dose Number Disparity, a dual-volume reformulation of the dose number (D_0), with Structural Properties Intelligence (SPI), a latent-variable projection of descriptors. Unlike principal component analysis, three-dimensional principal factor analysis (3D-PFA) separates shared from compound-specific variance; here, to our knowledge, it is first applied to developability. For five SGLT2 inhibitors (canagliflozin, empagliflozin, dapagliflozin, ertugliflozin and bexagliflozin), PFA resolved seven descriptors into three factors capturing 95.76% of common variance: dose–solubility disparity (F1, 59.44%), interfacial polar-surface energy (F2, 27.60%) and a minor residual factor (F3, 8.72%), with 4.24% descriptor-specific uniqueness. Dual-volume analysis revealed a compendial masking hazard: empagliflozin and bexagliflozin dissolve completely in 900 mL (D_0^Comp < 1) yet remain dissolution-limited at a physiological 250 mL (D_0^Clin > 1). Empagliflozin and ertugliflozin occupied the high-surface-energy pole of F2 (γs = 52.3–54.1 mN/m), consistent with non-wetting, clumping and floatation in pH 1.2 medium. Canagliflozin was an extreme F1 outlier owing to its 300 mg dose (D_0^Clin = 266.67). Influence diagnostics (Hotelling's T^2 limit 114.99; Q limit 1.044) confirmed an outlier-free model and identified empagliflozin as the most idiosyncratic member (uniqueness ≈ 0.95). A Quality-by-Design risk matrix links each compound to a preferred process path and mitigation.

Keywords: developability; dose number; biopharmaceutics classification; surface free energy; principal factor analysis; common-factor model; Quality by Design; SGLT2 inhibitors.

How to cite this article: Uthayathevan A, Sarkar A, Tharmaraj C, Jayaprakash NS., Predictive Developability of Gliflozins: Integrating Dose Number Disparity and Structural Properties Intelligence by Principal Factor Analysis. Int J Drug Deliv Technol. 2026;16(88s): 1730-1747; DOI: 10.25258/ijddt.16.88s.172



Graphical Abstract: The BCS III Developability Roadmap (Gliflozin's Case Study): From Molecular Risk to QbD Risk Mitigation

1. Introduction

The contemporary drug-discovery pipeline is exceptionally efficient at generating potent chemical matter, yet a large fraction of that matter is physically difficult to develop. A substantial majority of new small-molecule candidates emerging from combinatorial chemistry and high-throughput screening are poorly water-soluble, placing them into Class II or Class IV of the Biopharmaceutics Classification System (BCS) and into the high-risk quadrants of its developability-oriented successor, the Developability Classification System [1–3]. For these compounds, therapeutic performance is rarely dictated by equilibrium solubility in isolation. It emerges instead from the intersection of three quantities that are seldom evaluated together: the absolute mass that must dissolve to deliver a clinical dose, the fluid volume actually available in the relevant compartment, and the energetics of the crystal surfaces that must be wetted before dissolution can begin [4,5].

Historically, developability screening treated structural attributes and thermodynamic solubility as independent checkpoints cleared in sequence [6]. This compartmentalised logic is convenient but fragile. A compound with sub-microgram solubility can be entirely developable if its therapeutic dose is correspondingly small, whereas a molecule with moderate solubility can present intractable formulation problems when the clinical dose runs to hundreds of milligrams [7]. The same disconnect propagates downstream, where surface chemistry that is invisible to a solubility assay nonetheless governs powder flow, compaction and dissolution. The practical consequence is a recurring pattern of late-stage surprises: manufacturing failures during scale-up, unexpected bioavailability ceilings, and physically unstable formulations that perform acceptably in a quality-control vessel but disappoint *in vivo*. Moving beyond reactive, empirical formulation requires predictive protocols that extract physical intelligence from the simplest available inputs, structure and dose.

A central weakness of early developability workflows is their failure to couple clinical dose demand to localised fluid volume. The dimensionless dose number (D_o) introduced by Amidon and colleagues already encodes this coupling [2], yet in routine practice it is most often computed against a single, generous dissolution volume and then read as a pass/fail flag. We argue that the more informative quantity is the **difference** between dose numbers computed at two physically meaningful volumes. Compendial dissolution testing typically uses a 900 mL vessel, whereas the fasted human stomach offers on the order of 250 mL. A candidate can satisfy $D_o \leq 1$ in the laboratory while violating it in the body. To make this hazard explicit, the present work formalises the metric of **Dose Number Disparity**, contrasting a clinical dose number (D_0^{Clin}, 250 mL) with a compendial dose number (D_0^{Comp}, 900 mL). Where the two diverge across the unity threshold, a formulation may look promising under standard quality-control conditions yet fail in the restrictive fluid volumes encountered physiologically, a condition we term the compendial masking hazard.

Beyond the macroscopic dosing variables, a candidate's downstream behaviour is encoded in its crystal lattice and molecular topography. Descriptors such as topological polar surface area (TPSA), the estimated hydrophilic–lipophilic balance (HLB) and the crystal surface free energy (γs) jointly determine how a bulk powder interacts with processing equipment and aqueous media [8–12]. Individually, each descriptor is only weakly diagnostic; their value emerges when they are read together as a coordinated signature. We therefore introduce **Structural Properties Intelligence (SPI)**, a multidimensional approach that integrates molecular descriptors with macroscopic processing metrics through latent-variable projection.

Where most pharmaceutical chemometric workflows reach for principal component analysis (PCA), which decomposes the total variance of the data into composite axes [13–15], we instead adopt principal factor analysis (PFA), also termed principal axis factoring. PFA is a common-factor model that partitions the variance of each descriptor into a portion shared with other descriptors (the communality) and a descriptor-specific portion (the uniqueness) [16–19]. This distinction is not cosmetic. Because PFA models only shared covariance, the latent factors it returns represent the physical structure common to the descriptor set rather than mathematical composites inflated by descriptor-specific noise, and the residual uniqueness provides an explicit, interpretable measure of how atypical each candidate is. For a developability problem, where the goal is to identify the few latent physical dimensions that actually drive manufacturability, the common-factor formulation is conceptually better matched to the question than total-variance PCA.

To demonstrate the framework, we selected a homologous series of SGLT2 inhibitors: canagliflozin, empagliflozin, dapagliflozin, ertugliflozin and bexagliflozin. The class is unusually well suited to a comparative developability study because its members share a common C-glucoside core yet differ in defined structural appendages: a lipophilic thiophene in canagliflozin, a hydrophilic tetrahydrofuran ether in empagliflozin, a bridged dioxabicyclo skeleton in ertugliflozin, a low-polarity aglycone in dapagliflozin and an extended benzyl-ether arm in bexagliflozin. These modest substitutions translate into pronounced differences in polar surface footprint, surface energy and, importantly, clinical dose, which spans from 10 mg to 300 mg across the set [7]. The series thus offers a controlled gradient along which to observe how surface chemistry and dose loading ripple outward into wettability, agglomeration, punch filming and dissolution kinetics.

To the best of our knowledge, this study is the first to apply three-factor (3D) principal factor analysis, a latent common-variance method long established in the behavioural and analytical-chemistry literatures [16–23] but rarely, if ever, used for solid-state developability, to integrated pharmaceutical developability assessment, and the first to couple it with a dual-volume Dose Number Disparity metric and a Quality-by-Design (QbD) risk matrix across a homologous drug series. The prevailing multivariate tools in pharmaceutics are PCA and partial least squares regression [13–15,24,25]; the common-factor model has not, to our knowledge, been deployed to separate the shared biopharmaceutical structure of a candidate series from compound-specific idiosyncrasy in this way. We regard the priority claim as provisional pending a formal bibliometric search, but note that no prior report known to us integrates these three elements. The objectives of the study were therefore: (i) to formalise and apply the dual-volume Dose Number Disparity metric; (ii) to construct a 3D-PFA projection that resolves dose-driven and surface-driven risks onto orthogonal latent factors while quantifying descriptor uniqueness; (iii) to validate the resulting predictions against direct wettability of empagliflozin as qualitative observation; and (iv) to translate them into an actionable QbD risk matrix.

2. Materials and Methods

Dataset compilation

The assessment was built on a standardised biopharmaceutical descriptor matrix for the five gliflozins. Each compound was characterised by seven continuous descriptors spanning physicochemical and manufacturing domains: predicted equilibrium solubility (Cs, mg/mL); maximum single therapeutic dose strength (M, mg); clinical dose number (D_0^{Clin}, computed at 250 mL); compendial dose number (D_0^{Comp}, computed at 900 mL); topological polar surface area (TPSA, Å²); estimated hydrophilic–lipophilic balance (HLB); and predicted lattice surface free energy (γs, mN/m, numerically equivalent to mJ/m²). Solubility, HLB and surface-energy values were obtained as in silico predictions and should be read as model estimates rather than primary measurements; the dose strengths correspond to the highest marketed single-dose tablets. The complete baseline matrix is presented in Table 1.

Two derived flags were computed directly from the dose numbers. A compendial masking flag was raised whenever $D_0^{Comp} < 1 < D_0^{Clin}$, indicating that a compound would dissolve completely under compendial volumes but remain dissolution-limited at physiological volumes. A BCS class was assigned to each compound from its solubility and established permeability behaviour, yielding one Class IV member (canagliflozin) and four Class III members.

Table 1. Raw comprehensive biopharmaceutical property matrix for the gliflozin series. Cs, predicted equilibrium solubility; M, maximum single dose; D_0^{Comp} and D_0^{Clin} dose numbers at 900 mL and 250 mL; ΔD_o, disparity ratio; TPSA, topological polar surface area; HLB, estimated hydrophilic–lipophilic balance; γs, predicted lattice surface free energy.

Compound (BCS class)	Cs (mg/mL)	Max Dose (mg)	D_0^{Clin} (250 mL)	D_0^{Comp}, (900 mL)	ΔD_0 ratio	TPSA (Å²)	HLB	γs (mN/m)
Canagliflozin (IV)	0.0045	300	266.67	74.07	3.6	111.4	7.5	48.5
Empagliflozin (III)	0.030	25	3.33	0.93	3.6	115.6	9.2	52.3
Dapagliflozin (III)	0.173	10	0.23	0.064	3.6	106.4	8.8	46.8
Ertugliflozin (III)	0.119	15	0.50	0.14	3.6	120.2	9.5	54.1
Bexagliflozin (III)	0.0565	20	1.42	0.39	3.6	115.6	8.2	49.6

The Dose Number Disparity construct

The dose number is the dimensionless ratio of the concentration that a dose would generate if fully dispersed in the available fluid to the equilibrium solubility of the compound in that medium [2]:

$$D_0 = \frac{Maximum\ Daily\ Dose}{Predicted\ Intrinsic\ Solubility\ (C_s)\ x\ Volume\ of\ Dissolution\ Media(C_s)} \quad \ldots.(1)$$

where M is the maximum single dose (mg), V0 the available fluid volume (mL) and Cs the equilibrium solubility (mg/mL). Do ≤ 1 indicates that the available volume is thermodynamically sufficient to dissolve the entire dose, whereas Do > 1 signals a dissolution-limited, saturation-constrained system. We expanded this single-point metric

into a dual-volume analysis by computing the clinical dose number against a fasted-stomach volume and the compendial dose number against a standard United States Pharmacopeia (USP) Apparatus 2 vessel volume:

$$D_0 = \frac{Maximum\ Daily\ Dose}{C_s\ x\ Clinical\ Dissolution\ Volume\ (250\ mL)} \quad \text{....(2)}$$

$$D_0 = \frac{Maximum\ Daily\ Dose}{(C_s)\ x\ Compendial\ Dissolution\ Volume\ (900\ mL)} \quad \text{....(3)}$$

Because both expressions share M and Cs, their ratio collapses to a constant geometric factor, the Dose Number Disparity factor (ΔDo):

$$\Delta D_0 = \frac{Compendial\ Dissolution\ Volume}{Clinical\ Dissolution\ Volume} \quad \text{....(4)}$$

$$\Delta D_0 = \frac{900\ mL}{250\ mL} \quad \text{....(5)}$$

$$\Delta D_0 = 3.6 \quad \text{....(6)}$$

This fixed 3.60 ratio conceals an analytical inflection point. As a candidate's intrinsic properties slide along the solubility–dose continuum, it can enter a narrow band in which $D_0^{Comp} < 1$ and $D_0^{Clin} > 1$. Within this band a drug dissolves fully and appears unremarkable in a routine compendial test yet faces genuine dissolution limitation in the restricted volumes available *in vivo* [5]. Flagging this band is the principal diagnostic purpose of the disparity construct.

Interfacial energetics and thermodynamic wetting

The transition of a dry powder to a dispersed suspension is governed by the surface free energy of its crystal planes. Following Fowkes and the Owens–Wendt–Rabel–Kaelble treatment, the solid surface energy resolves into dispersive and polar components [8,9]:

$$\gamma s = \gamma s(d) + \gamma s(p) \quad \ldots.(7)$$

where γs(d) captures van der Waals and London dispersion forces and γs(p) captures hydrogen bonding and dipolar interactions. The wetting equilibrium at the solid–liquid–vapour line is described by Young's equation,

$$\gamma SV = \gamma SL + \gamma LV \cdot \cos\theta \quad \ldots.(8)$$

with θ the contact angle and γSV, γSL and γLV the solid–vapour, solid–liquid and liquid–vapour interfacial energies. The net tendency to wet is captured by the spreading coefficient S [10]:

$$S = \gamma SV - (\gamma SL + \gamma LV) = \gamma LV\,(\cos\theta - 1) \quad \ldots.(9)$$

Spontaneous wetting without mechanical assistance requires $S \geq 0$, equivalent to $\theta = 0°$. When the polar fraction of the exposed crystal faces is small, the solid–liquid interfacial energy against an aqueous phase remains high, S becomes negative, and the system actively rejects moisture, the physical origin of clumping and floatation [11].

Thermodynamics of particle agglomeration

The total surface free energy stored in a particulate mass scales with both its specific surface energy and its exposed area, $G_{surface} = \gamma s \cdot A_{total}$. Any spontaneous process reduces overall free energy ($\Delta G < 0$); for a powder with high γs the dominant route to lower energy is a reduction in exposed area through interparticle agglomeration [12]:

$$\Delta G = \gamma s \cdot \Delta A \quad \ldots.(10)$$

Because γs multiplies the area change, fine high-energy particles experience a strong thermodynamic driving force to coalesce ($\Delta A < 0$). This aggregation can cancel the dissolution advantage intended by micronisation, because the newly created internal interfaces are removed from contact with the bulk medium.

Principal factor analysis pipeline

To remove the scale incompatibility between descriptors (for example, Å² against mg), the raw 5 × 7 matrix was standardised by column to zero mean and unit variance (autoscaling):

$$X(scaled) = \frac{(X - \mu)}{\sigma} \quad \ldots.(11)$$

The autoscaled matrix was then decomposed by principal factor analysis rather than PCA. PFA expresses each standardised descriptor zj as a linear combination of a small set of latent common factors plus a descriptor-specific unique factor [16,18,19]:

$$zj = lj1 \cdot F1 + lj2 \cdot F2 + lj3 \cdot F3 + ej \quad \ldots.(12)$$

where lja are the factor loadings, Fa the latent common factors and ej the unique factor. Crucially, and unlike PCA, which eigen decomposes the full correlation matrix with a unit diagonal and therefore explains total variance [13,14], whereas PFA operates on a reduced correlation matrix in which the unit diagonal is replaced by communality estimates (h^2j, the variance of descriptor j shared with the others). The complementary quantity, the uniqueness $u^2j = 1 - h^2j$, is the descriptor-specific variance the common factors do not explain. Initial communalities were estimated as the squared multiple correlation of each descriptor on the remaining six and refined iteratively until convergence; loadings were obtained as the scaled eigenvectors of the converged reduced matrix [19]. The number of common factors was fixed at three on the basis of the scree criterion [21] and the eigenvalue pattern [22], following standard exploratory factor-analytic practice [23,24].

Two complementary diagnostics characterised model fit and influence. Hotelling's T^2 measured each sample's leverage, its distance from the model centre within the retained factor space [13]:

$$T^2(i) = \frac{\Sigma\, t^2(ia)}{\lambda(a)} \quad \ldots.(13)$$

with the 95% control limit derived from the F-distribution. Secondly, $T^2(limit) = [A(n-1)/(n-A)] \cdot F\alpha(A, n-A)$. In the common-factor setting the squared residual (Q) coincides with the sample uniqueness, the variance left in a sample once the common factors are removed:

$$Q(i) = \Sigma\,(X(ij,scaled) - \hat{X}(ij,scaled))^2 \quad \ldots.(14)$$

with its 95% limit approximated from a scaled chi-square distribution. Together, T^2 and the uniqueness coordinate define an influence space in which genuine high-leverage samples are separated from samples that are merely structurally atypical (high uniqueness).

In vitro *dissolution and wetting observation*

To validate the factor-analytic predictions, dissolution and wetting were observed directly on unformulated, high-energy raw powder. Tests used a USP Apparatus 2 (paddle) assembly with 900 mL of 0.1 N HCl simulated gastric fluid adjusted to pH 1.2, held at 37.0 ± 0.5 °C with paddle agitation of 50 rpm. A 25 mg powder charge, the highest comparable single-dose strength across the Class III members, was introduced onto the medium surface. High-resolution digital imaging at fixed intervals recorded macroscale wetting, dispersion kinetics, hydration pathways and any clumping or floatation at the air–liquid interface. These observations were interpreted against the spreading-coefficient and agglomeration relationships [26].

3. Results

Dose Number Disparity and the compendial masking hazard

The dual-volume analysis (Table 1, Figure 1) immediately stratified the series into three biopharmaceutical regimes. Dapagliflozin and ertugliflozin were comfortably dissolution-sufficient at both volumes D_0^{Clin} = 0.23 and 0.50; D_0^{Comp} = 0.064 and 0.14), reflecting their relatively high predicted solubility and low dose. Canagliflozin occupied the opposite extreme, dissolution-limited under every condition (D_0^{Clin}, = 266.67; D_0^{Comp} = 74.07), a direct consequence of its very low solubility (0.0045 mg/mL) combined with the highest dose in the class (300 mg). The diagnostically interesting compounds were empagliflozin and bexagliflozin, which fell squarely inside the compendial masking band: both satisfied $D_0^{Comp} < 1$ (0.93 and 0.39) yet violated $D_0^{Clin} > 1$ (3.33 and 1.42). A routine 900 mL dissolution test would therefore certify complete dissolution for these two molecules while a physiologically faithful 250 mL volume predicts saturation-limited release. The constant ΔD_o of 3.60 across all five compounds confirms that this hazard is a purely geometric property of the volume ratio, activated only when a candidate's intrinsic solubility–dose balance positions it across the unity threshold.

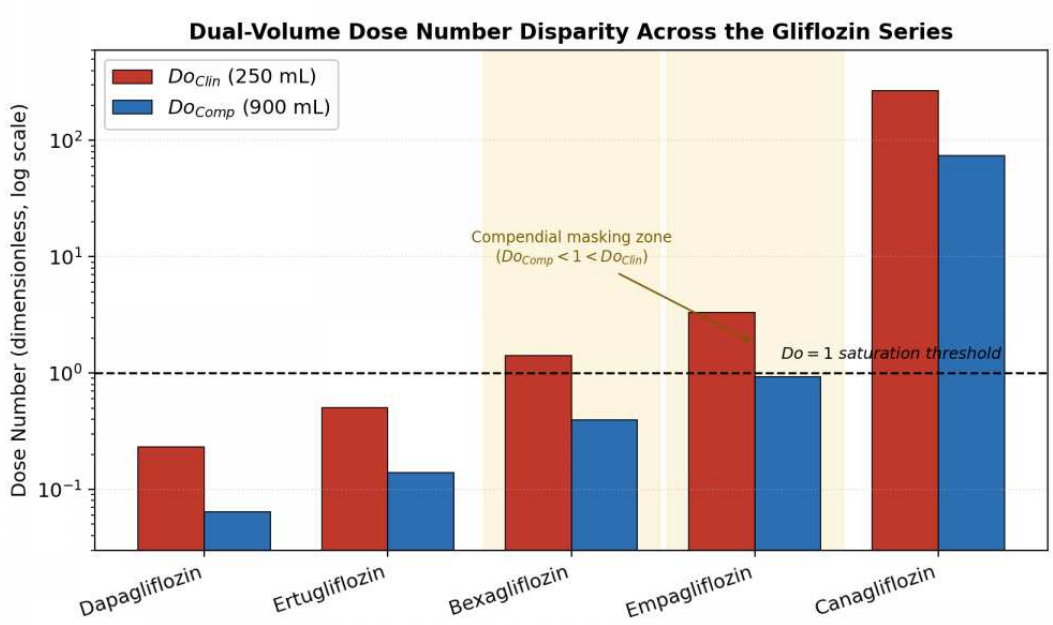


Figure 1. Dual-volume dose number disparity across the gliflozin series, plotted on a logarithmic scale. Clinical (250 mL) and compendial (900 mL) dose numbers are shown for each compound against the Do = 1 saturation threshold (dashed line). The shaded band marks the compendial masking zone, in which empagliflozin and bexagliflozin pass under compendial volumes ($D_0^{Comp} < 1$) yet fail at physiological volume ($D_0^{Clin} > 1$). Canagliflozin is solubility-limited under all conditions. Dapagliflozin and Ertugliflozin are highly soluble at the clinical dose.

Structure Property Intelligence

The pair-wise scatter plot matrix utilizes the Pearson correlation coefficient (r), a parametric statistic measuring the linear relationship between two continuous variables on a strict scale from -1.00 (perfect negative correlation) to +1.00 (perfect positive correlation) and a vlue close to 0 indicate no correlation. The Pearson correlation matrix shows an absolute positive correlation (r = 1.00) matching Maximum Daily Dose (mg) directly with both Clinical Dose Number ($\boldsymbol{D_0^{Clin}}$) and Compendial Dose Number ($\boldsymbol{D_0^{Comp}}$). Conversely, Predicted Aqueous Solubility (mg/mL) displays a strong negative correlation against administered dose mass (r = -0.62) and an identical inverse relationship (r = -0.59) with both dose numbers. Looking at structural properties, Total Polar Surface Area (TPSA, $Å^2$) exhibits a robust, direct positive correlation (r = 0.93) with Predicted Surface Energy (mN/m). In contrast, electronic and surface boundary properties like TPSA (r = -0.06) and Surface Energy (r = -0.15) show negligible, weak correlations when paired directly against absolute clinical or compendial dousing metrics across the gliflozin dataset.

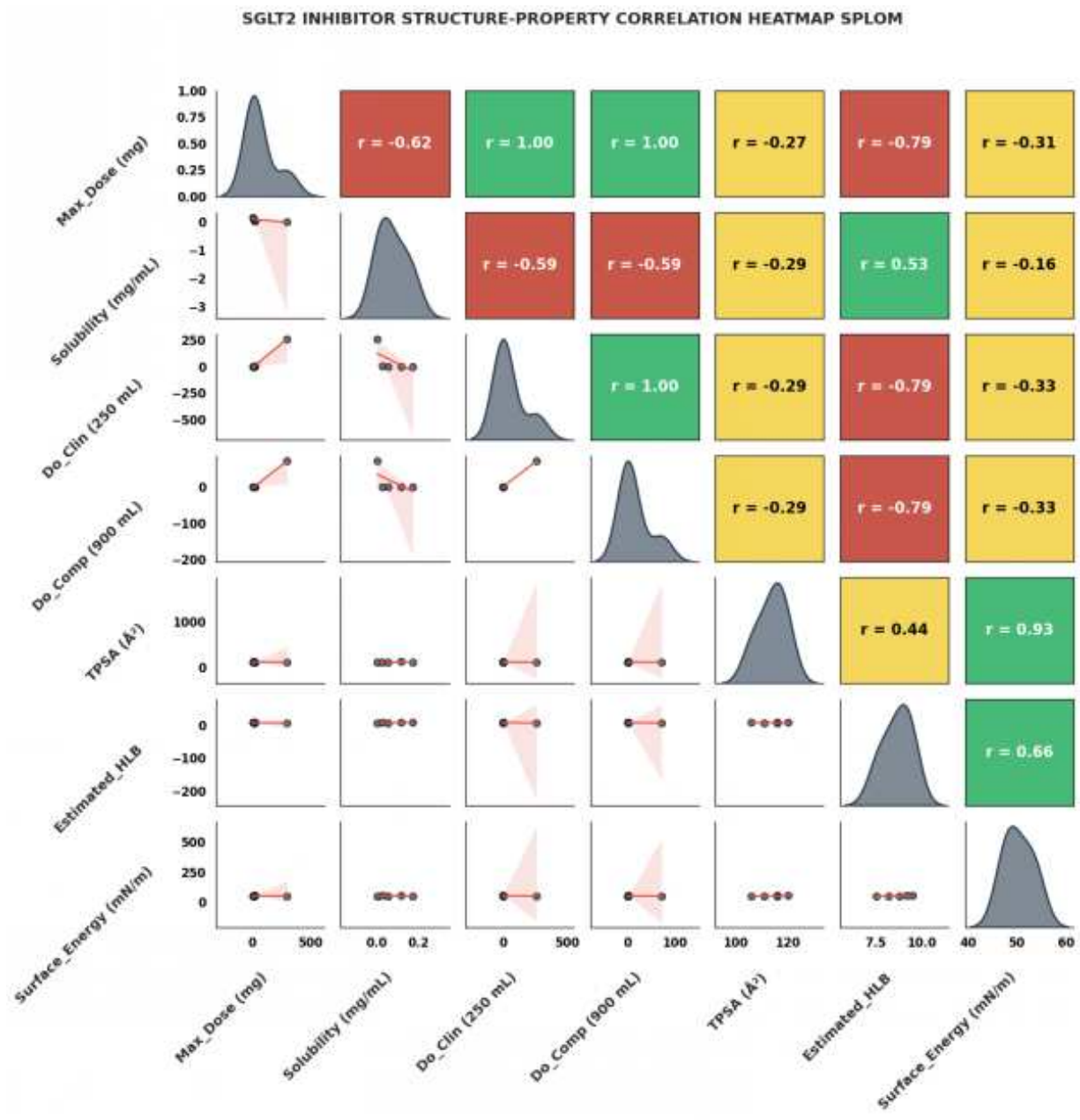


Figure 2. SGLT2 Inhibitor Structure Property Correlation Heatmap SPLOM

Common-variance apportionment by Principal Factor Analysis

Unlike principal component analysis (PCA), which maximizes total variance (including error/noise), principal factor analysis (PFA) separates the shared common variance from unique/error variance. It iteratively targets the communalities along the diagonal of the correlation matrix, making it a much more rigorous approach for uncovering hidden physical chemistry and biopharmaceutical relationships. PFA of the auto-scaled matrix extracted three interpretable common factors that together accounted for 95.76% of the common variance, distributed as F1 = 59.44%, F2 = 27.60% and F3 = 8.72% (Figure 3A-C, common-variance breakdown). The residual 4.24% represents descriptor-specific uniqueness, variance not shared across the descriptor set and therefore not attributable to any latent physical factor. This partition is the central methodological advantage of the common-factor approach over PCA: rather than forcing the data to 100% explained variance, PFA explicitly quarantines the idiosyncratic component, so the retained factors describe only the structure genuinely shared across descriptors. The dominance of the first two factors (87.04% jointly) indicates that gliflozin developability is, to first order, a two-dimensional problem, consistent with the intrinsic dimensionality of a five-sample dataset.

Factor loadings: orthogonal separation of dose and surface domains

The loading structure (Figure 3A-C, molecular factor loadings) revealed a clean and physically meaningful partition. F1 was defined almost entirely by the bulk dosing and solubility descriptors: maximum dose, D_0^{Clin} and D_0^{Comp} loaded near +1.0, while predicted solubility (≈ −0.60) and estimated HLB (≈ −0.80) loaded negative. F1 therefore functions as an interpretable **dose–solubility disparity factor**, with high positive scores denoting compounds whose absolute dose overwhelms their solubility. F2 was driven by the surface descriptors, predicted surface energy

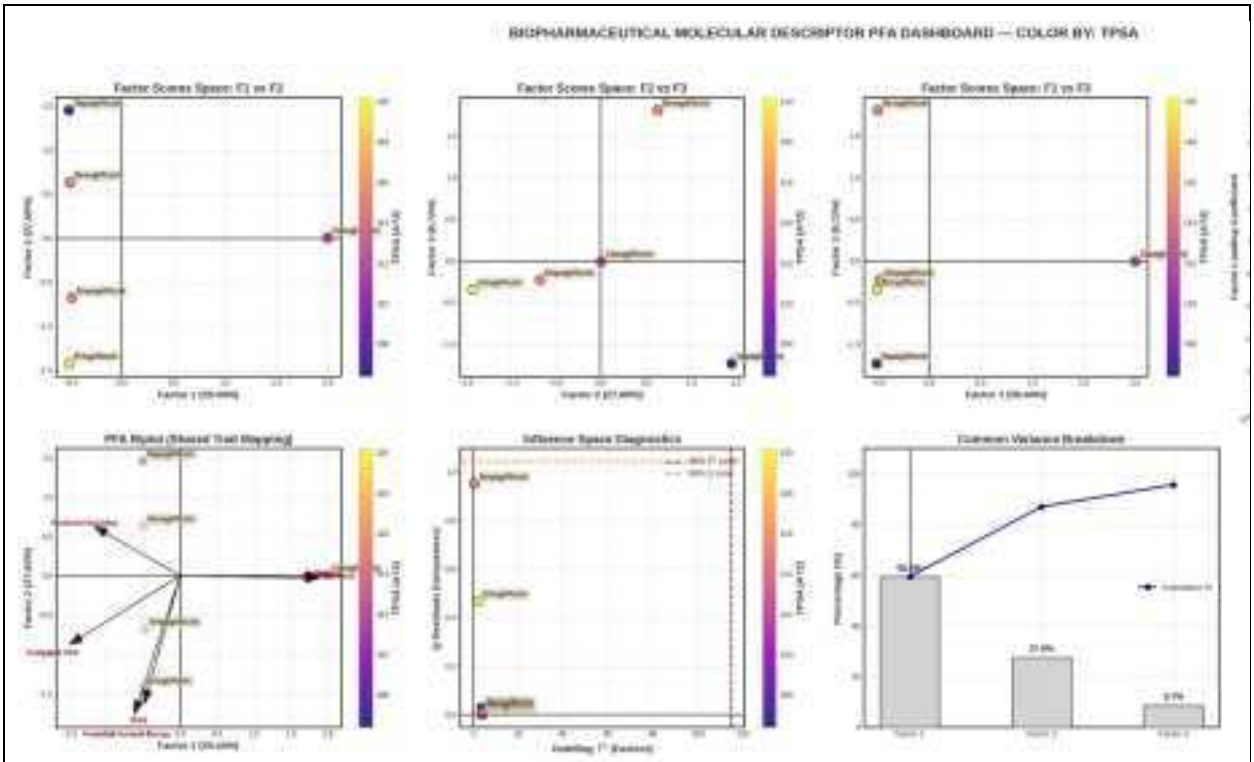


Figure 3A. Three-dimensional principal factor analysis (3D-PFA) dashboards for the gliflozin descriptor matrix, with factor score maps (F1×F2, F2×F3, F1×F3), molecular factor loadings, the shared-trait PFA biplot, influence-space diagnostics (Hotelling's T² versus as a function of TPSA (Å²)

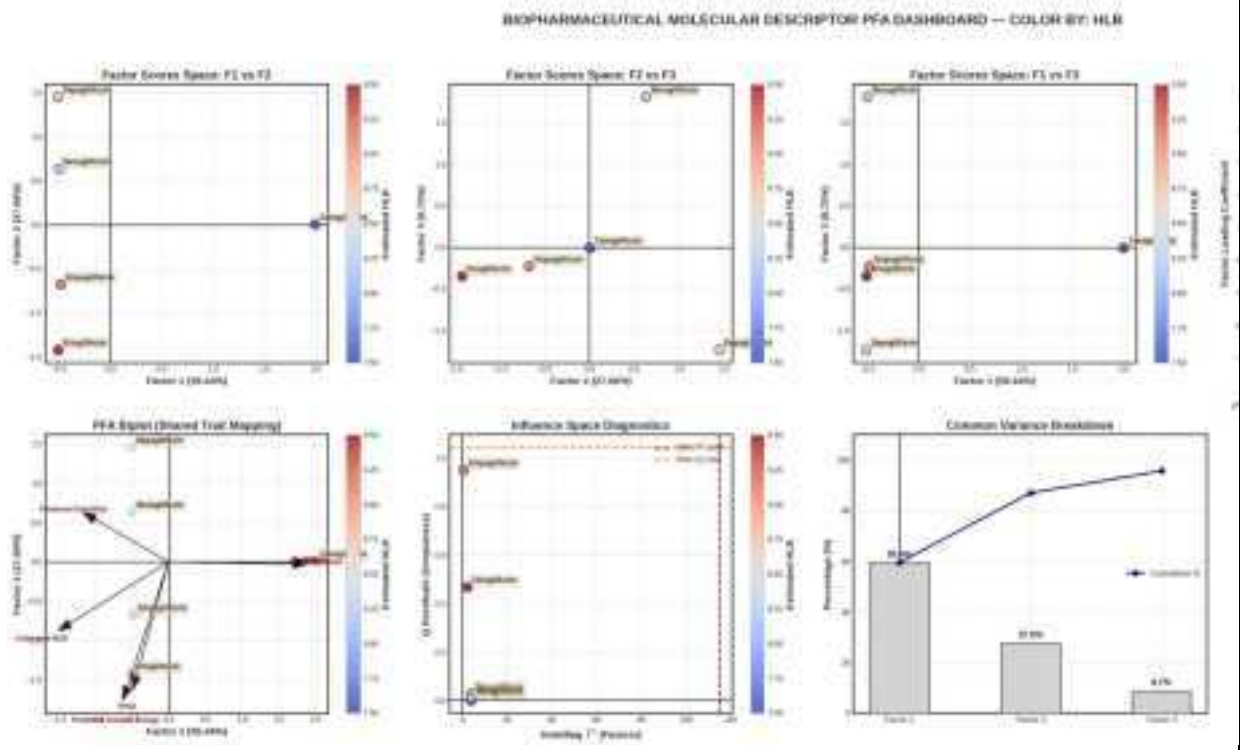


Figure 3B. Three-dimensional principal factor analysis (3D-PFA) dashboards for the gliflozin descriptor matrix, with factor score maps (F1×F2, F2×F3, F1×F3), molecular factor loadings, the shared-trait PFA biplot, influence-space diagnostics (Hotelling's T² versus as a function of HLB

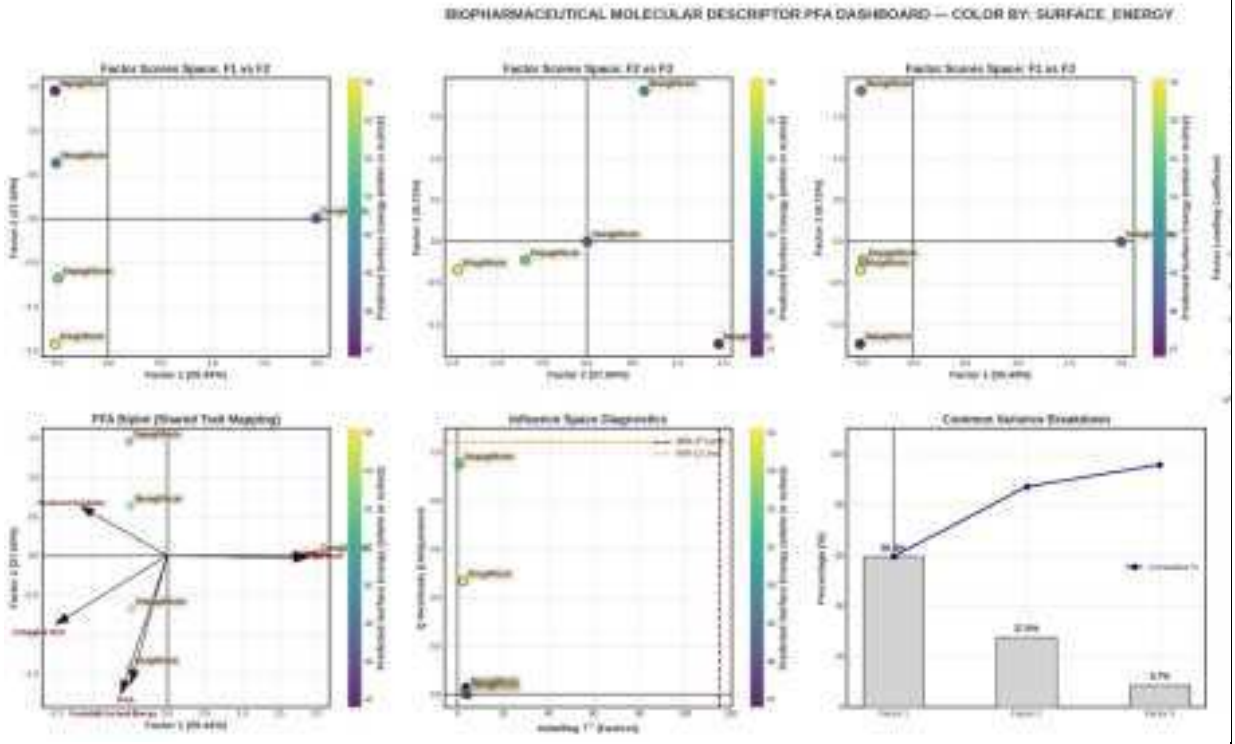


Figure 3C. Three-dimensional principal factor analysis (3D-PFA) dashboards for the gliflozin descriptor matrix, with factor score maps (F1×F2, F2×F3, F1×F3), molecular factor loadings, the shared-trait PFA biplot, influence-space diagnostics (Hotelling's T² versus as a function of Predicted Surface Energy (mN/m)

(≈ −0.92) and TPSA (≈ −0.87) loaded strongly, with HLB (≈ −0.43) contributing, making it an **interfacial polar-energetics factor**. Because the sign of a factor is arbitrary, the negative loadings simply mean that high-surface-energy, high-TPSA compounds occupy the negative pole of F2. The minor F3 captured a residual polar/solubility nuance (TPSA ≈ +0.42, solubility ≈ −0.52). The near-orthogonality of F1 and F2 is the central structural result: the absolute dose burden of a compound and the polar energetics of its crystal surface are statistically independent latent dimensions of risk and must be managed by different formulation levers. summarise the variance capture (F1 59.44%, F2 27.60%, F3 8.72%; cumulative 95.76%) and 95% control limits (Hotelling's T^2 114.99; Q/uniqueness 1.044).

Factor-score clustering and structural mapping

Projecting the compounds into the factor-score space reproduced the structural logic of the series. Along the F1 axis, canagliflozin separated decisively to the far positive edge (F1 ≈ +2.0), an extreme dosing outlier produced by its 300 mg dose, an order of magnitude above the 10–25 mg of its congeners, which drives D_0^{clin} to 266.67. The remaining four compounds clustered near the negative-F1 region, distinguished instead by their F2 coordinate. At the negative-F2 (high-surface-energy) pole sat ertugliflozin (F2 ≈ −1.45; γs = 54.1 mN/m, TPSA = 120.2 Å²) and empagliflozin (F2 ≈ −0.65; γs = 52.3 mN/m, TPSA = 115.6 Å²), consistent with the highly oxygenated bridged dioxabicyclo skeleton of ertugliflozin and the hydrophilic tetrahydrofuran ether core of empagliflozin. At the opposite, positive-F2 pole sat dapagliflozin (F2 ≈ +1.47; γs = 46.8 mN/m, TPSA = 106.4 Å²), whose surface is dominated by dispersive van der Waals interactions and the smallest polar footprint in the set, with bexagliflozin at an intermediate positive position. Each compound is thus located at the intersection of a dose-risk coordinate (F1) and a surface-risk coordinate (F2), allowing the dominant liability of each molecule to be read directly from its position.

Influence diagnostics and the uniqueness coordinate

To confirm that the projection was not distorted by a single dominant sample, all five compounds were mapped into the Hotelling T^2–uniqueness influence space (Figure 3A-C, influence-space panels). The 95% control limits were T^2 = 114.99 and Q/uniqueness = 1.044. Every gliflozin plotted far within the T^2 boundary, confirming that no molecule, not even the dosing outlier canagliflozin exerts disproportionate leverage on the model alignment. In the uniqueness dimension, empagliflozin carried by far the highest value (≈ 0.95), projecting close to but beneath the 1.044 limit, with ertugliflozin next (≈ 0.47) and the remaining three near zero. The elevated uniqueness of empagliflozin is the common-factor model's explicit signal that its tetrahydrofuran ether configuration contributes variance not shared with the rest of the descriptor set, a structural idiosyncrasy that a total-variance PCA would silently fold into its components. Because the value remains within the control boundary, the model is statistically valid and outlier-free, while correctly identifying empagliflozin as the most atypical member of the class.

Experimental validation: wetting and dissolution behaviour

Direct dissolution observation in pH 1.2 medium confirmed the risk predicted at the negative-F2, high-surface-energy pole (Figure 3). Rather than sinking and dispersing, the bulk powder immediately formed cohesive macro-aggregates that entrained air and floated at the liquid–gas interface. This non-wetting response is the macroscopic expression of an unfavourable spreading coefficient ($S < 0$) and a high solid–liquid contact angle, driven by the dispersive character of the unformulated crystal habit [10,11]. The spontaneous formation of floating clumps minimised the area exposed to the medium ($\Delta A \ll 0$) and established a stagnant, locally saturated boundary layer within the core of each aggregate, severely restricting mass transfer in accordance with the Noyes–Whitney relationship [26]. The experiment confirms that the surface energies resolved along F2 are not abstract coordinates but direct determinants of bulk wetting and dissolution behaviour.



Figure 4. Macroscale wetting behaviour of unformulated high-energy gliflozin powder on simulated gastric fluid (0.1 N HCl, pH 1.2; USP Apparatus 2, 50 rpm, 37 °C). Across top-down (left, centre) and side (right) views, the powder resists wetting, forms cohesive

4. Discussion

A mechanistic continuum from surface energy to dissolution failure

The convergence of the factor-analytic projection and the dissolution experiment supports a single mechanistic narrative that runs from molecular topography to clinical performance. Compounds at the negative-F2 pole expose crystal faces rich in polar functional groups and carry elevated surface free energy (52.3–54.1 mN/m for empagliflozin and ertugliflozin). Counter-intuitively, this high lattice energy does not promote aqueous wetting; against a polar medium the unformulated dispersive faces still present a large solid–liquid interfacial energy, so the spreading coefficient is negative and the powder rejects water [10,11]. The same high γs simultaneously provides the thermodynamic driving force for interparticle agglomeration, because coalescence reduces exposed area and therefore total surface free energy [12]. The resulting floating clumps trap air, collapse the effective dissolution area and erect a saturated boundary layer, so that mass transfer is throttled exactly where the dose number already predicts difficulty. Surface energetics (F2) and dose–solubility burden (F1) thus reinforce one another at the point of dissolution, even though they are statistically independent latent factors.

Downstream manufacturing liabilities

The functional groups responsible for non-wetting in the dissolution vessel create parallel liabilities on the production line, and the negative-F2 compounds are predicted to be the most troublesome. Strong interparticle cohesion produces poor powder flow, electrostatic charging and bridging in feed hoppers, a particular concern for ertugliflozin and empagliflozin. During high-speed compression, the same high-energy polar faces preferentially adhere to stainless-steel tooling rather than bonding cohesively within the compact, manifesting as punch filming, picking and core lamination that degrade content uniformity and tablet integrity [27]. Canagliflozin presents the inverse problem: its liability is not surface cohesion but mass. Because the 300 mg active dominates the tablet, mechanical behaviour during compaction is governed by the API's own crystal habit rather than by excipients, raising the risk of capping and lamination unless the drug load is engineered explicitly [27]. Reading F1 and F2 together therefore predicts not only dissolution risk but the specific unit operations most likely to fail.

Translation to a Quality-by-Design risk framework

To convert these mechanistic insights into manufacturing controls, the factor coordinates were mapped onto a Quality-by-Design (QbD) risk matrix that links each compound's position in the SPI space to critical material attributes, critical quality attributes and a preferred process path (Table 2), in keeping with regulatory expectations for science- and risk-based development [28,29]. The matrix ranks blend-flow, punch-filming and wetting/floatation risks for each molecule and pairs them with targeted controls. Ertugliflozin, at the negative-F2 extreme, is rated critical on all three axes; empagliflozin, also at the negative-F2 pole and additionally inside the masking band, is rated high-to-critical; canagliflozin, the extreme F1 outlier, carries a high-drug-load process risk but only moderate surface risk; dapagliflozin, at the positive-F2 (low-energy) pole, is the most benign; and bexagliflozin sits at intermediate risk on every axis. The value of the matrix is that it assigns risk before development begins, using only descriptors available at candidate selection.

Table 2. Quality-by-Design critical-quality-attribute risk-mitigation matrix for the gliflozin series, ranking processing risks by position within the 3D-PFA projection and linking each compound to a preferred process path and recommended formulation controls. F2 sign follows the loading convention of Figure 2, in which the negative pole denotes high surface energy.

Compound	F1 status	F2 status	Flow risk	Filming risk	Wetting risk	Preferred process	Recommended controls
Canagliflozin	Extreme +	Near-zero	Low	Low	Moderate	High-shear wet granulation	Surfactant (e.g., SLS) to lower interfacial tension; elevated binder to carry high drug load.
Empagliflozin	Slight –	Negative (high γs)	High	High	Critical	Fluid-bed wet granulation	Hydrophilic polymer masking (HPMC/PVP) over high-energy planes; colloidal SiO_2 glidant.
Dapagliflozin	Slight –	Positive (low γs)	Low	Low	Low	Direct compression / roller compaction	Standard lubrication (Mg-stearate); compatible with continuous manufacturing.
Ertugliflozin	Slight –	Negative extreme (max γs)	Critical	Critical	Critical	Wet granulation / particle modification	Silica pre-blend to lower surface charge; anti-adherent punch coating (e.g., CrN).
Bexagliflozin	Slight –	Positive (intermediate)	Moderate	Moderate	Moderate	Roller compaction / dry granulation	Optimised extra-granular glidant to prevent flow stuttering.

Risk-specific formulation mitigation

The matrix points to two distinct mitigation philosophies aligned with the two orthogonal risk factors. For the negative-F2 cohesion problem (ertugliflozin and empagliflozin), direct compression is strongly discouraged because of punch filming and poor flow. These molecules are better served by wet granulation, ideally fluid-bed granulation, which deposits a hydrophilic polymer layer such as hydroxypropyl methylcellulose or povidone across the high-energy crystal faces. This coating masks the raw cohesive forces, lowers the solid–liquid interfacial tension and restores uniform dispersion during dissolution [11,12], while colloidal silica can be added to support flow. For the high-F1 dosing problem (canagliflozin), the formulation must instead manage drug load. Because the API dominates the compact, high-shear wet granulation is preferred, incorporating a surfactant such as sodium lauryl sulfate to counter the hydrophobic thiophene ring and facilitate wetting [7], with sufficient binder to preserve mechanical integrity at high drug fraction. Dapagliflozin, lowest on both risk axes, is amenable to direct compression or roller compaction with conventional lubrication, and bexagliflozin's balanced profile suits dry granulation with an optimised extra-granular glidant. The framework thus does not merely flag risk; it prescribes the lever appropriate to each risk's physical origin.

Why a common-factor model, and comparison with conventional screening

Principal Factor Analysis (PFA) was employed to identify the latent common-cause developability factors governing dissolution performance and manufacturability, distinguishing shared variance from variable-specific noise [16]. Within the proposed Quality by Design (QbD) framework, this (shared variance) enables early recognition of inherent, API-related formulation risks like surface wettability, dissolution sensitivity, lubrication susceptibility, and process dependency, thereby, providing a rational basis for formulation design, process selection, and control-strategy development. By contrast, special-cause variation arising from manufacturing excursions, equipment variability, raw-material changes, or operational deviations is managed through lifecycle monitoring, continued process verification, and risk-based controls consistent with ICH Q10 principles.

The value of separating these two variance sources is clear against conventional triage, which treats solubility, dose, and surface descriptors as independent thresholds and, in several cases, returns a misleading verdict. A single 900 mL dissolution screen would certify empagliflozin and bexagliflozin as unproblematic, masking their physiological dissolution limitation; an isolated solubility cut-off would group ertugliflozin with benign compounds despite its peak surface energy and severe non-wetting; and a dose-only filter would flag canagliflozin without distinguishing its dosing problem from the surface problem of the high-energy pair. The integrated PFA approach corrects all three errors simultaneously. Crucially, the choice of PFA over the more familiar PCA is consequential rather than stylistic: by modelling only shared variance, PFA returns latent factors that map cleanly onto interpretable physical mechanisms—a dose factor and a surface factor—and yields an explicit uniqueness coordinate that isolates structurally atypical candidates such as empagliflozin [16,19,23]. A total-variance PCA would instead absorb that idiosyncratic variance, reporting 100% explained variance and obscuring precisely the signal a developability scientist needs. The common-factor formulation is therefore better matched to identifying the few latent dimensions that drive manufacturability [27].

Limitations and future directions

Several limitations temper these conclusions and define the next steps. First, the analysis rests on five compounds and seven descriptors; with such a small sample the intrinsic dimensionality is bounded and the communality estimates, control limits and loadings, while internally consistent, should be regarded as hypothesis-generating rather than definitive [23,24]. Expanding the matrix to a larger and more chemically diverse compound set, and bootstrapping the factor solution, would sharpen the statistical limits and test the stability of the dose and surface factors. Second, solubility, HLB and surface energy were treated as in silico predictions; replacing them with measured biorelevant solubilities, inverse gas chromatography surface energies and experimental contact angles would convert the projection from a predictive screen into a calibrated one [30–32]. Third, the dissolution validation was qualitative and conducted on unformulated powder at a single pH; quantitative release profiling of the masking-band compounds at 250 mL versus 900 mL, and of formulated prototypes, would directly test the central biopharmaceutical claim. Fourth, the priority claim for 3D-PFA in developability is provisional and should be substantiated by a formal bibliometric search before publication. None of these limitations undercuts the core finding that dose disparity and surface energetics are separable, jointly necessary latent dimensions of developability; they instead define a clear route to a validated, generalisable tool.

5. Conclusion

This study shows that developability assessment is substantially improved by replacing isolated, single-point screens with an integrated framework built on Dose Number Disparity and Structural Properties Intelligence, resolved for the first time, to our knowledge, by three-dimensional principal factor analysis. Applied to a homologous series of SGLT2 inhibitors, the common-factor model compressed a seven-descriptor space into three interpretable latent factors

capturing 95.76% of the shared variance, while quarantining 4.24% as descriptor-specific uniqueness. The dominant dose–solubility disparity factor (F1) isolated absolute dosing burden and identified canagliflozin as an extreme high-load problem requiring robust high-shear granulation. The interfacial polar-energy factor (F2) mapped a clean gradient of surface excess energy and moisture affinity, with empagliflozin and ertugliflozin anchoring the high-risk pole for clumping, equipment adhesion and poor wetting, a prediction confirmed by direct dissolution observation in acidic media. The dual-volume dose-number analysis additionally exposed a compendial masking hazard for empagliflozin and bexagliflozin, compounds that pass a standard 900 mL test yet remain dissolution-limited at physiological volume, and the uniqueness coordinate flagged empagliflozin as the structurally most atypical member. Consolidated into a QbD risk matrix, these findings give formulators a proactive, target-specific basis for selecting process paths and mitigations before scale-up, illustrating a shift from empirical formulation toward a predictive, design-driven developability science.

Declarations

Funding. No Funding Received for this Work.

Author Contributions (CRediT)

Anbalagan Uthayathevan: Conceptualization, Formal Analysis, Methodology, Software, Data Curation, Data Interpretation, Visualization, Writing – Original Draft. **Chozharajan Tharmaraj:** Formal Analysis, Software, Data Curation, Data Interpretation, Visualization, Writing – Original Draft. **Jayaprakash N S:** Data Interpretation, Visualization, Writing – Original Draft. **Dr Ashish Sarkar:** Project Administration, Supervision, Writing – Review & Editing.

Acknowledgements

The authors acknowledge YBN University, Ranchi and apex laboratories private limited, Chennai for providing the necessary facilities and support to conduct the research. We also extend our gratitude to the the laboratory staff for their technical assistance.

Declaration of Generative AI Use

During the preparation of this work, the authors used ChatGPT, Claude, Gemini, and NotebookLM to support literature understanding and concept summarization. All outputs were critically reviewed, verified, and substantially edited by the authors. The authors take full responsibility for the content of this article.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Due to confidentiality reasons, the primary data are not publicly available. The corresponding authors may be contacted for further information.

References

1. Williams HD, Trevaskis NL, Charman SA, Shanker RM, Charman WN, Pouton CW, et al. Strategies to address low drug solubility in discovery and development. Pharmacol Rev. Elsevier; 2013;65:315–499.
2. Amidon GL, Lennernäs H, Shah VP, Crison JR. A theoretical basis for a biopharmaceutic drug classification: the correlation of in vitro drug product dissolution and in vivo bioavailability. Pharm Res. Springer; 1995;12:413–20.
3. Butler JM, Dressman JB. The developability classification system: application of biopharmaceutics concepts to formulation development. J Pharm Sci. Elsevier; 2010;99:4940–54.
4. Lipinski CA. Rule of five in 2015 and beyond: Target and ligand structural limitations, ligand chemistry structure and drug discovery project decisions. Adv Drug Deliv Rev. Elsevier; 2016;101:34–41.
5. Hörter D, Dressman JB. Influence of physicochemical properties on dissolution of drugs in the gastrointestinal tract. Adv Drug Deliv Rev. Elsevier; 2001;46:75–87.
6. Dahan A, Miller JM, Amidon GL. Prediction of solubility and permeability class membership: provisional BCS classification of the world's top oral drugs. AAPS J. Springer; 2009;11:740–6.
7. Kawakami K. Modification of physicochemical characteristics of active pharmaceutical ingredients and application of supersaturatable dosage forms for improving bioavailability of poorly absorbed drugs. Adv Drug Deliv Rev. Elsevier; 2012;64:480–95.
8. Fowkes FM. Attractive forces at interfaces. Ind \& Eng Chem. ACS Publications; 1964;56:40–52.
9. Owens DK, Wendt RC. Estimation of the surface free energy of polymers. J Appl Polym Sci. Wiley Online Library; 1969;13:1741–7.
10. Buckton G. Contact angle, adsorption and wettability—a review with respect to powders. Powder Technol. Elsevier; 1990;61:237–49.
11. Hancock BC, York P, Rowe RC. The use of solubility parameters in pharmaceutical dosage form design. Int J Pharm. Elsevier; 1997;148:1–21.

12. Rowe RC. Binder-substrate interactions in granulation: a theoretical approach based on surface free energy and polarity. Int J Pharm. Elsevier; 1989;52:149–54.
13. Hotelling H, others. The generalization of Student's ratio. Springer; 1931;
14. Wold S, Esbensen K, Geladi P. Principal component analysis, Chemometr. Intell. Lab., 2, 37--52. 1987.
15. Bro R, Smilde AK. Principal component analysis. Anal methods. Royal society of chemistry; 2014;6:2812–31.
16. Malinowski ER, Howery DG. Factor analysis in chemistry. Wiley New York; 2002.
17. Spearman C. " General Intelligence" Objectively Determined and Measured. Appleton-Century-Crofts; 1961;
18. Thurstone LL. Multiple factor analysis. Psychol Rev. Psychological Review Company; 1931;38:406.
19. Harman HH. Modern factor analysis. Univ Chicago. 1976;
20. Brereton RG. Chemometrics: data analysis for the laboratory and chemical plant. John Wiley \& Sons; 2003.
21. Cattell RB. The scree test for the number of factors. Multivariate Behav Res. Taylor \& Francis; 1966;1:245–76.
22. Kaiser HF. The application of electronic computers to factor analysis. Educ Psychol Meas. Sage Publications Sage CA: Thousand Oaks, CA; 1960;20:141–51.
23. Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ. Evaluating the use of exploratory factor analysis in psychological research. Psychol Methods. American Psychological Association; 1999;4:272.
24. Ab C. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Pr assess, res eval. 2005;10:1–9.
25. Massart DL, Vandeginste BG, Buydens LMC, Lewi PJ, Smeyers-Verbeke J, Jong S De. Handbook of chemometrics and qualimetrics. Elsevier Science Inc.; 1998.
26. Noyes AA, Whitney WR. The rate of solution of solid substances in their own solutions. J Am Chem Soc. ACS Publications; 1897;19:930–4.
27. Sun CC. Materials Science Tetrahedron—A Useful Tool for Pharmaceutical Research and Development. J Pharm Sci [Internet]. Elsevier; 2009 [cited 2025 Dec 15];98:1671–87. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0022354916329616
28. Yu LX, Amidon G, Khan MA, Hoag SW, Polli J, Raju GK, et al. Understanding pharmaceutical quality by design. AAPS J. Springer; 2014;16:771–83.
29. ICH EMA. Q8 (R2) Pharmaceutical Development—Scientific Guideline.
30. Newell HE, Buckton G, Butler DA, Thielmann F, Williams DR. The use of inverse phase gas chromatography to measure the surface energy of crystalline, amorphous, and recently milled lactose. Pharm Res. Springer; 2001;18:662–6.
31. Heng JYY, Bismarck A, Lee AF, Wilson K, Williams DR. Anisotropic surface energetics and wettability of macroscopic form I paracetamol crystals. Langmuir. ACS Publications; 2006;22:2760–9.
32. Jolliffe I. Principal Component Analysis. Int Encycl Stat Sci. Springer; 2025. p. 1945–8.