RSC Advances

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *RSC Adv.*, 2026, **16**, 10179

Dimensionality reduction of COSMO-RS molecular descriptor using functional principal component analysis (FPCA) for organic solvent mapping

Luis Eduardo Ramirez Cardenas,*[a] Rachid Ouaret,[b] Vincent Gerbaud,[b] Ivonne Rodriguez Donis[a] and Sophie Thiebaud-Roux*[a]

Within the context of a transition towards greener and safer solvents, we describe a framework facilitating solvent screening. Traditional approaches rely on experimental solubility data or computational methods such as COSMO-RS. In parallel, similarity maps can be helpful to explore alternative molecules similar to working solvents. For developing solvent maps, Principal Component Analysis (PCA) offers limited applicability when dealing with complex molecular descriptors such as the σ-potential derived from COSMO-RS theory. In this study, we propose the application of Functional Principal Component Analysis (FPCA) as a more suitable dimensionality reduction technique for solvent mapping, leveraging the functional nature of σ-potentials. A database of 1588 solvents was analyzed, extending previous reported datasets with the inclusion of industrially relevant and sustainable candidates. FPCA enables a two-dimensional representation of the solvent space with minimal information loss (0.5%), directly associating the principal components with electron donor and acceptor characteristics. In this space, solvent clustering naturally emerges, facilitating the identification of structurally and functionally similar solvents. Three case studies are presented to illustrate the practical implications of the approach. Overall, this methodology provides a suitable framework for solvent substitution, whether as a preliminary screening step or as a part of computer-aided solvent design tools, contributing to more sustainable chemical practices.

Received 27th October 2025
Accepted 29th January 2026

DOI: 10.1039/d5ra08246c

rsc.li/rsc-advances

Open Access Article. Published on 19 February 2026. Downloaded on 6/19/2026 1:24:24 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Introduction

Solvents are mainly considered as liquid chemicals that can dissolve or extract other compounds without altering chemically either the compounds or themselves. However, reactions, separation, formulation, and cleaning processes generally require a significant excess of solvents to achieve the desired conversion rate, the product purity and recovery yield. The rise of the petrochemical industry has dramatically increased the availability and diversity of solvents. The global solvent market reaches approximately 28 million tons annually, with the paint, coating, cleaning and pharmaceutical industries accounting for the majority of consumption.[1,2] Despite their widespread use, many traditional solvents are derived from non-renewable sources and pose significant environmental and health risks due to their toxicity, volatility, flammability, and poor biodegradability. The European Union's REACH regulation aims to mitigate these risks by restricting or banning hazardous chemicals. Several commonly used solvents, such as benzene, toluene and dichloromethane, are already regulated under REACH,[3] which continues to identify other substances likely to face future restrictions. In particular, compounds such as amides, nitroaromatics, certain hydroxyethers and chlorinated solvents fall within the scope of these regulatory measures. At present, many existing chemical processes continue to rely on harmful and toxic solvents, due to their low cost and wide availability (*e.g.*, hexane, methanol, *N*,*N*-dimethylformamide, toluene, dichloromethane). However, stricter regulations are pushing industries to look for safer and more sustainable alternatives. The transition towards greener solvents is essential for aligning industrial practices with environmental and safety standards. Industry generally prioritizes the reuse of existing substances, whereas academia concentrates on developing more efficient approaches to identify alternative solvents. Several in-house solvent selection guides have been developed by the pharmaceutical industry primarily focused on safety, health, and environmental (EHS) impacts with limited consideration of the required physical properties. The CHEM21 consortium[4] reviewed major guides from GSK,[5–7] AstraZeneca,[8] Sanofi,[9] ACS GCI-PR,[10] Rowan University,[11] and ETH Zürich,[12] ultimately creating a consensus guide within the pharmaceutical sector, which can be seen reflected in tools like GreenScore.[13] Byrne *et al.*[14] reviewed existing solvent guides and

[a]*Univ. Toulouse, Toulouse INP, INRAE, Laboratoire de Chimie Agro-Industrielle (LCA), Toulouse, France. E-mail: luiseduardo.ramirezcardenas@toulouse-inp.fr*
[b]*Univ. Toulouse, CNRS, Toulouse INP, Laboratoire de Génie Chimique (LGC), Toulouse, France*

© 2026 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 19 February 2026. Downloaded on 6/19/2026 1:24:24 PM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

concluded that traditional formats have reached their limitations, highlighting the need for more systematic approaches tailored to assist non-expert users in solvent selection.

There exist many other techniques to find alternative solvents including (1) comparing solvents in basis of their required properties or descriptors similarity, (2) through experimental trial and error, or with (3) Computer-Aided Molecular Design (CAMD). Solvent selection typically involves solving a multi-criteria optimization problem, where factors related to key functionality in the process (solubility, phase transition temperatures) must be balanced with other practical factors such as chemical stability, transport properties (viscosity, density, surface tension), energy-related properties (phase change enthalpy, specific heat), and economic factors like cost and availability. In addition to these functional and practical properties, environmental, health and safety (EHS) criteria are becoming increasingly important in the selection process. These factors have been considered to map similarity between solvents before coupled with dimensionality reduction and clustering techniques[8,15–18] to decrease correlation between these factors and outline base behaviors that may arise when considering large dimensional spaces.

In this study we apply Functional Principal Component Analysis (FPCA),[19,20] to create an effective mapping, suitable to the functional nature curve-like descriptors obtained with the Conductor Like Screening MOdel for Real Solvents (COSMO-RS).[21] Our study encompasses 1588 molecules (liquid at room temperature), expanding on the list established by Moity *et al.*[18] by incorporating additional liquid compounds from the COSMOtherm database, green solvents suggested by Clark *et al.*,[22] and new alternative solvents selected based on in-house data. Ultimately, the effectiveness of this method as a tool for supporting solvent substitution with the required solubilizing properties is assessed using three case studies from different application domains. This methodology could be integrated as a preliminary step in computer-aided solvent conception tools to identify groups of closely similar solvents before the application molecular design.

Background

Among the crucial aspects of the solvent's role is its solvation capacity. It is primarily driven by solute–solvent interactions, which are broadly classified into two categories: non-specific interactions, such as van der Waals forces and ion–dipole attractions, and specific interactions, including hydrogen bonding, electron-pair donor–acceptor interactions, and solvophobic effects.[23] The traditional "like dissolves like" principle in solvent selection provided access to empirical models and was first formalized through Hildebrand's solubility parameter (δ)[24] which is based on cohesive energy density. Hansen[25] later refined this concept by introducing a more nuanced model with three distinct parameters: dispersion (δ_d), dipolar (δp), and hydrogen bonding (δ_H) interactions, offering an improved understanding of solute–solvent compatibility. The closer these parameters are between a solvent and solute, the more likely they are to dissolve. Kamlet–Abboud–Taft[26] (KAT) formalized the solvation process by dividing dissolution into three steps: cavity formation, solute separation, and energy release from interactions, leading to solvatochromic parameters (dipolarity π^*, hydrogen bond accepting ability β, and hydrogen bond donating ability α) for solvents and solutes. Jin *et al.*[27] proposed a 10-step framework to identify and develop bio-based alternatives to conventional solvents based on Hansen solubility and KAT parameters.

Efforts to transition directly from the chemical structure and chemical group interactions to molecular properties go back to the 1940s, when one of the earliest semi-empirical methods, the group contribution (GC), was developed.[28] Since then, advances in property prediction methodologies have been driven by the need to move beyond traditional trial-and-error approaches for solvent substitution towards solvent selection tools. GC methods are the most commonly used semi-empirical models as Quantitative Structure–Property Relationships (QSPR) in CAMD because they offer a straightforward way to estimate pure compound properties based on the contributions of individual structural chemical groups. Hukkerikar *et al.*[29] updated and improved the parameters of 18 properties of the GC$^+$ models, combining both group-contribution and atom connectivity index methods and using a large experimental data-set. Within this set of models, certain approaches enable the estimation of both the Hildebrand solubility parameter and the Hansen solubility parameters. The prediction of environmental, health, and safety (EHS) properties is commonly based on Quantitative Structure–Activity Relationships (QSAR). Molecular descriptors, commonly including physicochemical characteristics, structural features, and electronic attributes are correlated with the EHS properties using experimental data.[30] Software platforms like VEGA-QSAR[31] exemplify this approach by bundling over 90 QSAR models, enabling *in silico* prediction of toxicological, ecotoxicological, environmental and physico chemical properties without additional empirical data. However, a major limitation of both GC and QSAR methods is their dependence on quality and diversity of experimental datasets, which often lacks sufficient chemical diversity, leaving many parameters unavailable and thus limiting the design space in CAMD applications.

Full predictive theoretical methods have advanced rapidly since the early 21st century, offering a means to compare solvents without relying on experimental data. Early approaches relied on *ab initio* and density functional theory (DFT) calculations to describe the electronic structure of molecules, providing fundamental insights into solute–solvent interactions. However, the direct application of these methods to bulk liquids was limited by their high computational cost and the complexity of capturing collective solvent effects. To overcome these challenges, hybrid models such as continuum solvation frameworks were introduced, representing the solvent as a dielectric medium surrounding the solute. Building on this foundation, more sophisticated approaches like the Conductor-like Screening Model (COSMO)[32] and its extension, COSMO-RS,[21] incorporated statistical thermodynamics to connect quantum chemical data with macroscopic properties, enabling accurate predictions of solubility, activity coefficients, and

 © 2026 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 19 February 2026. Downloaded on 6/19/2026 1:24:24 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

partitioning behaviour. These advances have established quantum chemistry–based models as powerful tools for rational solvent design and the exploration of environmentally benign alternatives.

In the context of solvent substitution, several molecular descriptors are often taken into consideration, obstructing the search for alternative molecules, due to computational cost, and information overlap. Dimensionality reduction techniques such as PCA are frequently combined with clustering techniques to reduce redundant information, indicate similarity and guide the selection or design of substituent solvents. Chastrette[15] pioneered a solvent classification system using a multiparametric statistical approach considering 83 substances. This approach is based on six experimental physical properties (the Kirkwood function (K), molecular refraction (MR), molecular dipole moment (μ), the δ parameter of Hildebrand, index refraction (n), boiling point (bp)) along with the HOMO and LUMO predicted energies. PCA was employed to reduce the original eight-dimensional space to a three-dimensional space. Since the first decade of the 21st century, several data-driven tools have been developed integrating experimental data with PCA for solvent selection and substitution. Launched in 2009 as part of a European industry-academic collaboration, the SOLVSAFE tool applied PCA to a dataset of 347 molecules encompassing 11 chemical families and utilizing 52 structural descriptors. Integrating PCA results with predicted toxicity and ecotoxicity profiles allowed the identification of safer solvent alternatives.[33] Similarly, both the AstraZeneca[8] and Syngenta[34] applied PCA to create simplified "maps" of solvent space, enabling visual representation and comparison of solvents through multivariate statistical analysis of their properties. AstraZeneca's tool[8] assessed 272 solvents, representing a wide range of chemical types, based on seven normalized EHS criteria while the Syngenta tool[34] allowed the users to select parameters for identifying potential solvents from 209 molecules. Katritzky *et al.*[16] developed QSPR models to predict 127 polarity scales using 168 theoretical descriptors. These descriptors reflect key intermolecular interactions involved in solvation, including cavity formation, electrostatic polarization, dispersion forces, and hydrogen bonding.

Descriptors derived from COSMO-RS, called σ-moments, were first proposed by Klamt and Eckert.[35] This approach transforms the σ-potential into a Taylor-series expansion with respect to the surface charge density σ. The coefficients obtained from this expansion capture individual physical information. M_0 represents the total molecular area, M_1 the negative total charge, M_2 the electrostatic energy, and M_3 the skewness or asymmetry of the profile.[36] This methodology allows for an expansion up to 6 components, although standard applications truncate this series up to 3 or 4 components. Hydrogen bonding behaviour is treated separately based on specific thresholds to define donor and acceptor regions.[37] This approach has already been tested for property prediction by several authors.[35,36,38]

A novel solvent classification approach was proposed by Durand *et al.*,[17] based solely on molecular structure and COSMO-RS.[21] Expanding on earlier work by Chastrette *et al.*,[15] the method aimed for a solvent classification from theoretical descriptors by analysing the σ-potential curves of 153 compounds using PCA. Solvents were then classified into ten families using k-means. Following the 3D representation, Moity *et al.*[18] mapped 138 sustainable solvents across the ten solvent classes, based on the principle that solvents with σ-potential curves are theoretically capable of dissolving the same solute in the absence of ionic interactions. The study highlighted the potential of this method as a tool for guiding solvent replacement and the design of new solvents with the desired solubilizing capabilities. However, PCA, the dimensionality reduction technique used for the σ-potential curves, is not well-suited to this type of data. The resulting 3D representation is difficult to interpret, as solvent classes often overlap and are poorly distinguished along the third dimension, with a low information load compared to the other two more important dimensions. These limitations hinder the effective identification of suitable replacement solvents.

Overview of applied techniques

Understanding solvation phenomena and solvent–solute interactions is critical in the rational design of sustainable solvents. Among the computational tools developed for this purpose, COSMO-RS has emerged as a robust method, providing a quantum chemistry-based framework to estimate thermodynamic properties of mixtures. Several implementations of the theory are currently available, including the commercial packages COSMOtherm[39] and ADF-COSMO,[40] as well as alternative formulations such as COSMO-SAC[41] and the open-source platform openCOSMO-RS.[42] Together, these implementations provide diverse computational environments through which COSMO-RS methodologies can be applied across both academic and industrial contexts.

Additionally, FPCA offers a powerful statistical technique for extracting dominant patterns from functional data such as the sigma–potential curves generated by COSMO-RS. An overview of each technique is presented below.

COSMO-RS fundamental principles

COSMO[32] is a solvation model for calculating the electrostatic interaction between solute and solvent. It is based on the quantum chemical description of molecules in a scaled conductor medium, forming a cavity around the substance. The molecular surface is discretized into tiny segments (Fig. 1). Each segment is characterized by its area, position, and surface charge density, which assumes that the medium around the molecule is a perfect conductor ($\varepsilon \rightarrow \infty$). The polarization mirrors the medium's response to the solute's electrostatic potential. After determining the surface polarization charges, COSMO calculates the solvation free energy by evaluating the interaction between the induced surface charges and the electrostatic potential of the solute, and scaling it according to the solvent's dielectric constant.

While COSMO provides an initial estimate of solvation free energy by modeling the solute in a conductor-like medium, COSMO-RS refines these predictions by incorporating a statistical thermodynamic framework that accounts for real solvent

Open Access Article. Published on 19 February 2026. Downloaded on 6/19/2026 1:24:24 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

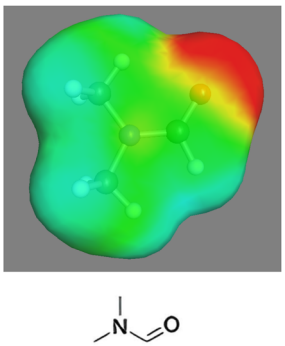

Fig. 1 Molecular structure and COSMO σ-surface of dimethylformamide.

effects and molecular interactions. COSMO-RS integrates the quantum chemical COSMO calculations of individual molecules with a statistical thermodynamic treatment of their pairwise interactions. Fig. 1 shows the surface charge distribution of dimethylformamide (DMF) from COSMO calculations, subsequently converted into a 'σ-profile' (Fig. 2), a histogram of surface polarization charge densities across the molecule. The geometrical information from the *ab initio* computations is dropped, since only segment–segment interactions matter from a thermodynamic point of view.

The σ-potential describes the interaction energy between segments of surface charge density σ and the molecule. The σ-potential curve encodes the electrostatic information of a molecule, effectively representing the affinity of the solvent for a surface charge σ. Because the calculations required for the σ-potential are purely *ab initio*, no experimental data are required. Another advantage is that all molecules share the same reference state of a conducting medium, which can later be scaled to more accurately represent the dielectric properties of the real environment surrounding the molecule. This makes the σ-potential a suitable candidate as a universal molecular descriptor. However, it does not capture all molecular information. Combinatorial contributions arising from the molecule relative volume and area are considered in COSMO-RS as separate contributions during the calculation of thermodynamic phase equilibria.

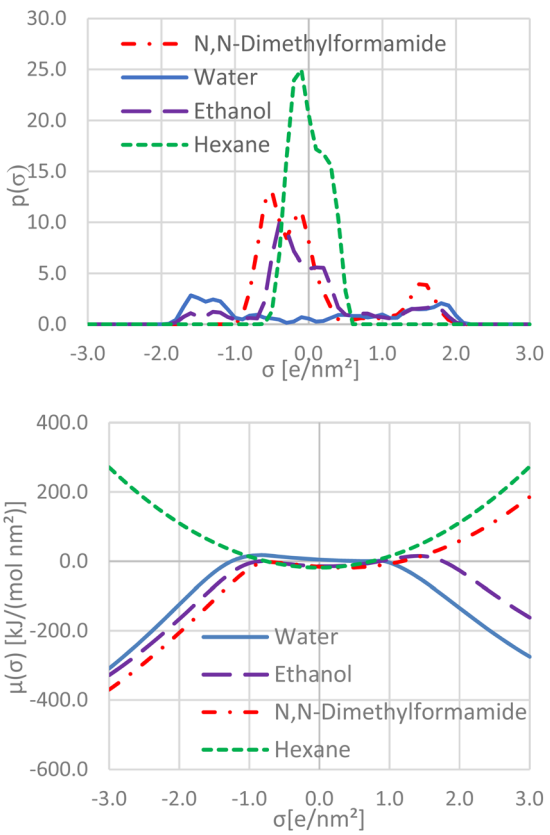

Fig. 2 σ-Profile (up) and σ-potential (down) of several representative solvents.

Fig. 2 illustrates the σ-profiles and σ-potentials of four different molecules. Apolar solvents like hexane are characterized by centered σ-profiles describing large neutral charge surfaces, and U-shaped σ-potentials where minimal interaction energy can be found only for neutral σ (-1 e nm$^{-2} \leq \sigma \leq 1$ e nm^{-2}). Polar amphiprotic solvents are also fairly symmetrical but with many segments in the most polar zones of the σ-profile, and with σ-potentials characterized by *m*-shaped curves where the interaction energies reach the lowest points at the extremes of σ values indicating a very low interaction capacity with apolar solvents and solutes. Aprotic polar solvents are asymmetrical in both their σ-profiles, with many segments with high positive σ values, and σ-potentials having "S" shaped curves with the best interactions occurring for negative σ values ($\sigma \leq -1$ e nm^{-2}). The behavior of the solvent can be inferred from the shape of the σ-potential. Further treatment of the σ-potential with reduction dimentionality techniques can improve the interpretation of the meaning behind this descriptor.

Dimensionality reduction

The classification and comparison of solvents often require the use of multiple descriptors, which introduces complexity in both visualization and interpretation. When multiple physicochemical properties are considered simultaneously, direct comparisons between solvents become difficult, particularly when the data reside in a high-dimensional space. However, the use of such descriptors frequently leads to redundancy either by capturing overlapping information (such as molecular mass, surface and volume, or polarity through dipole moment, and dielectric constant), or through the interdependence of physicochemical properties (boiling point, vapor pressure, enthalpy of vaporization), which are often derived from the same molecular characteristics. Additionally, the methods used to calculate these descriptors might share a common origin, whether from thermodynamics, quantum chemistry, or spectroscopy, further compounding the redundancy. To address this, PCA is often applied to decorrelate the descriptors, extract dominant patterns and facilitate the visualization of solvents in a reduced space.

Indeed, PCA is a statistical technique commonly used to reduce the dimensionality of complex datasets containing correlated variables, while preserving as much variability as

© 2026 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 19 February 2026. Downloaded on 6/19/2026 1:24:24 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

possible. It works by transforming the original set of variables into a smaller set of uncorrelated variables, known as principal components (PCs), which are linear combinations of the original variables. The first principal component captures the maximum variance in the data, with each subsequent component accounting for the largest remaining variance, while being orthogonal to the preceding components.

Although PCA is well suited for classical molecular descriptors expressed as discrete variables the situation considered by Durand *et al.*[17] departs from this standard framework. Their analysis relies on the σ-potential, which is inherently functional rather than discrete. Treating such a curve as a long vector of sampled values allows PCA to be applied formally, but it overlooks the structural nature of the data. In practice, this creates several methodological limitations. Functional observations contain strongly correlated values across the domain, which artificially inflates dimensionality and exposes PCA to issues such as sparsity, unstable loadings, and overfitting. Moreover, the discretization required to convert a function into a vector introduces an arbitrary dependence on the sampling resolution; different grids may lead to different principal components and, consequently, to different interpretations. These limitations highlight that conventional PCA is not theoretically aligned with the continuous nature of σ-potentials, and motivate the use of an approach that explicitly accounts for the functional structure of the data.

After applying PCA, results similar to those found by Durand *et al.*[17] were obtained. Fig. 3 shows that a 98.6% of the variance is captured using the first 5 principal components of these results. Indeed, we found that the 98.8% of the variance was represented with 5 components. The first two principal components obtained by PCA were tied to the electron donor and acceptor character of the molecule accounting for the majority of the variance in the data, approximately 75% (Fig. 3). The third PC was attributed to the lipophilicity of the molecules. However, as the lipophilic character of a molecule is primarily determined by the central region of the σ-potential curve, where variation is minimal, the differentiation captured by this principal component has a limited effect. This can be seen in the clustering of solvents in the proposed classification by Durand *et al.*[17] where several families are overlapped due to the third dimension. The remaining components were not tied to other meaningful physical property.

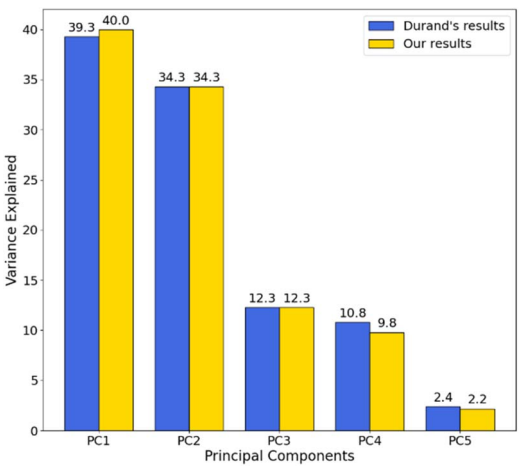

Fig. 3 Information on the Principal Components obtained with PCA by Durand *et al.*[17]

A key difference separating classical Principal Component Analysis (PCA) from its functional counterpart (FPCA) rests in both their mathematical foundations and the type of data they are meant to handle. Classical PCA works with finite-dimensional vectors, where each observation is a fixed set of discrete measurements. Its objective is to find linear combinations of these variables that capture the largest possible share of the total variance. This method inherently treats the data as a collection of isolated points, with no assumed relationship, such as continuity or smoothness between one measurement and the next, meaning that the underlying process generating the data lacks abrupt, irregular, or jagged changes.

FPCA, on the other hand, originates from functional data analysis. In this case, each observation is treated as a smooth function defined over a continuous domain, such as time, wavelength, or, as in our case, the σ-potential. Instead of computing a covariance matrix for discrete variables, FPCA estimates the covariance structure across the entire functional space and performs a spectral decomposition. The result is a set of smooth, orthonormal eigenfunctions that describe the main modes of variation in the data. Unlike PCA loadings, which assign weights to separate data points, FPCA eigenfunctions represent coherent functional patterns such as overall shifts, changes in shape, or smooth deformations across the domain.

This distinction becomes crucial when analysing data that are inherently continuous. Take the example of σ-potential: these are not simply lists of independent descriptors, but smooth curves representing the distribution of molecular surface charge density across the σ-scale. Their physical interpretation relies on the shape and smoothness of σ-potential, *i.e.*, adjacent values reflect gradual shifts in local polarity, and the overall curve conveys meaningful information about a molecule's electrostatic character. For applying classical PCA to a discretized version of a σ-potential ignores this functional nature, treating neighbouring σ-values as statistically independent and thereby overlooking the continuity that gives the profile its chemical relevance. In such cases, apparent "patterns" may arise as artifacts of the discretization grid rather than as genuine physicochemical features. (For example, for sigma between 2.999 and 3 [e nm^{-2}], the potential variation is very slight).

FPCA is designed to respect the functional form of σ-potentials. In fact, FPCA is an extension of PCA to situations where data consist of functions, not vectors. It explicitly models smoothness, reduces noise from discretization or measurement, and produces a low-dimensional representation grounded in chemically interpretable variations. For example, the leading functional principal components might correspond to a systematic shift in polarity towards more positive or negative σ-regions, a broadening or narrowing of the central charge distribution, or a change in the balance between polar and non-polar surface regions. These modes align with established

Open Access Article. Published on 19 February 2026. Downloaded on 6/19/2026 1:24:24 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

physicochemical principles and offer a more intuitive framework for characterizing solvent diversity.

Therefore, when the aim is to reduce dimensionality while retaining the intrinsic continuous structure of σ-potentials, FPCA provides a theoretically sound and scientifically coherent alternative to classical PCA. It bridges statistical methodology with the physical reality of molecular surface properties, making it not merely a technical choice, but a conceptually appropriate one for this type of data.

Methods

The diagram in Fig. 4 illustrates the overall methodology developed in this study to construct a solvent mapping framework based on the functional representation of σ-potentials. The workflow begins with a database of 1588 solvents, integrating literature-reported compounds together with additional industrially relevant and sustainable candidates. This dataset feeds into a preprocessing stage, where σ-potentials are computed using COSMO-RS and harmonized to ensure consistency across the full set of molecules. The preprocessed σ-potentials are then passed to a functional representation module, which treats each σ-potential as a smooth continuous curve. This step is essential, as it preserves the intrinsic continuity and physicochemical meaning of the σ-scale, avoiding the artifacts and information inflation associated with vectorized discretization. Outputs from FPCA analyses converge in a solvent-mapping module, which constructs a two-dimensional representation of the solvent space. The addition of new molecules or mixtures into the mapping then only requires their addition into the initial database to obtain their σ-potential before the application of FPCA.

Data set

The dataset of liquid molecules was compounded from the following sets: the database proposed by Gramatica *et al.*[43] containing 153 organic solvents, the COSMOtherm2024 DATABASE,[39] which was filtered to count only molecules in liquid state at normal temperature, the green solvents studied by Moity *et al.*,[18] and a list of greener solvents to which other molecules compiled in our laboratory have been added. The final dataset contains 1588 molecules and can be found in the SI.

COSMO-RS calculations

Molecules not contained within the COSMObase 2024 underwent a conformational analysis using the COSMOconf2024 script using the in-built BP-TZVP parametrization. The most relevant conformers go through an *ab initio* Density Functional Theory (DFT) calculation in Turbomole.[44] DFT/COSMO geometry optimizations were performed according to the standard method for COSMO-RS using the B88-PW86 functional with a triple zeta valence polarized basis set (TZVP).[45] COSMO surfaces, σ-profiles, and σ-potentials were generated *via* the software COSMOtherm 2024 using the TZVP parametrization. The relative contributions of each conformer are determined from their COSMO energy and chemical potential using a Boltzmann distribution. The σ-potential describes the chemical potential of a surface segment with a given screening charge density (σ) in a specific solvent. The σ-potentials are given for charge densities between −3 to 3 e nm^{-2} for 61 evenly distributed points, creating a matrix for the database of solvents of 1588 * 61.

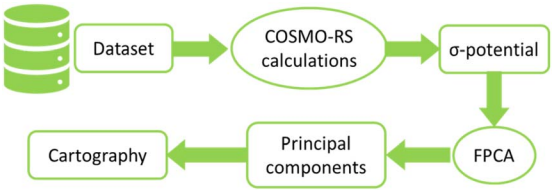

Fig. 4 Diagram of the steps for the methodology.

FPCA implementation

FPCA was applied on a database of 1587 organic compounds and water. The new descriptors generated by the dimensionality reduction are related to the physical properties of the molecules. The new space permits an easy visualization of the related properties. The results were compared with previous efforts to represent and classify solvents through the use of dimensional reduction and clustering techniques, while its experimental validation is explored through the solvency comparison of several molecules widely reported in the literature.

FPCA calculations were performed on the standardized values of the database matrix of σ-potentials in python software using the scikit-fda,[46] a package offering support for Functional Data Analysis (FDA). The shape of the functional components as well as the coordinates of each solvent in the new FPCA space were recuperated for further analysis.

Solubilization case studies

In order to validate our solvent classification, known solvents for several solutes were represented and identified in the FPCA space. The nitrocellulose example of solute proposed by Hansen[47] and analyzed by Chastrette *et al.*,[15] and Durand *et al.*,[17] was studied. Solvents generated by Moity *et al.*[18] were also considered. The other cases studied were the data of solubility of Ester Gum,[47] and triolein.[48]

FPCA results and discussion

Applying FPCA to the dataset of molecules yielded two significant outcomes: the visualization of the average curve showing the effect of adding or subtracting the principal components, and the placement of each solvent in the two-dimensional space defined by the variations of both principal components.

FPCA represents the principal components as the positive and negative effect on the average curve from the dataset. Fig. 5 illustrates the effect of each Principal Component (PC), with the first component having a high dependency of the electronic charge density (ECD) on electron-donating sites (δ^-), which can be related to the hydrogen bond accepting capacity. Because

© 2026 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 19 February 2026. Downloaded on 6/19/2026 1:24:24 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

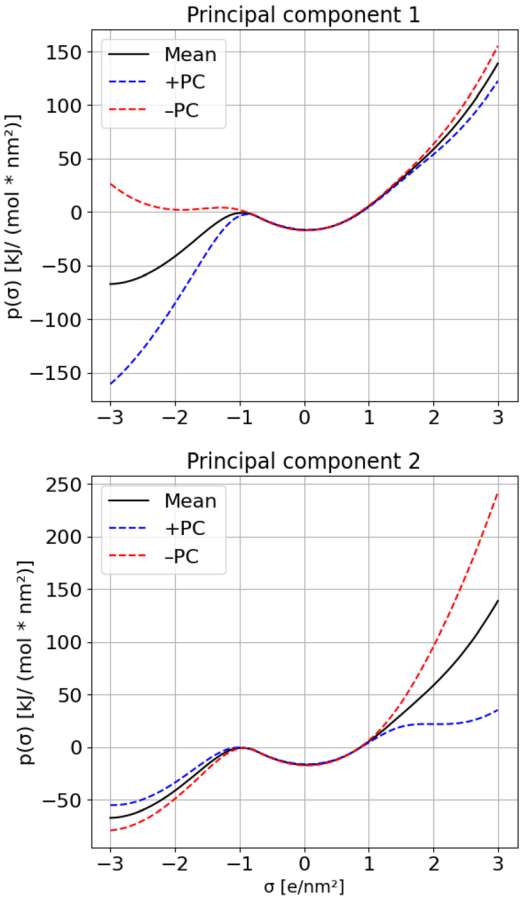

Fig. 5 Mean σ-potential (black), positive (red), and negative (blue) effect of the Principal Components obtained with FPCA.

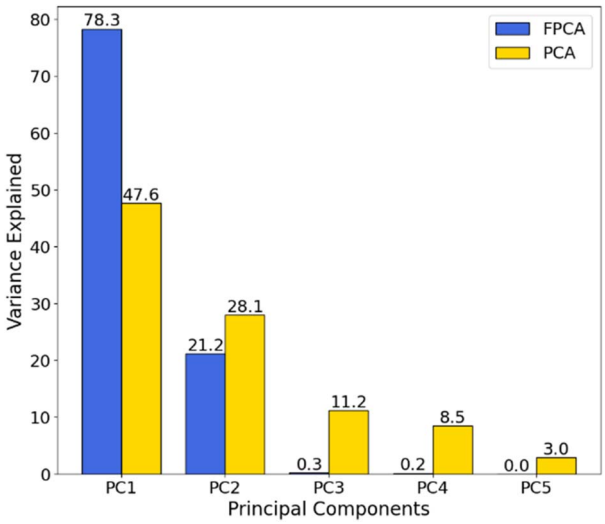

Fig. 6 Variance capture by the Principal components obtained with FPCA and compared against PCA for our organic solvent database.

most of the solvents in our database are aprotic ranging from low to high polarity, the largest percentage lies in the first PC associated with the ECD. The second PC reflects on the electron-deficient sites (δ^+) and the positive charge density (PCD), can then be associated with the Brønsted and Lewis acidity character of the solvents.

Considering only the first two PC results in a minimal information loss of 0.5% (Fig. 6). Applying PCA to the same dataset results in a 98.4% of the variance being captured by the first five principal components, which is consistent with the findings reported by Durand *et al.*[17] There is an increase on the first component suggesting that there are more solvents with high ECD yet the meaning of the PC remain the same.

After the application of FPCA molecules can be represented in a simple two-dimensional plane, facilitating easy interpretation (Fig. 7). As solvents are located further to the left their ECD increases, whereas moving to the right it gradually decreases until it disappears completely. On the other hand, the acidity increases downwards and diminishes upwards. Consequently, groups of solvents can be identified based on their relative positions in the FPCA plane. Apolar solvents, like hexane, toluene, benzene, are found in the upper right corner, with the lowest ability to either give or accept electrons. Aprotic polar solvents, such as dioxane, pyridine and DMF, are located in the upper section of the plane with an increasing ECD as they move closer to the left. Protic compounds capable of donating a proton to form a hydrogen bond are placed in the lower section of the plane with an increasing acid character towards the lower zones in the *y*-axis, and can be divided in two groups: hydrogen bond acceptor such as alcohols, water, and formamide, and poor acceptor such as phenol and trifluoracetic acid.

The use of only two components to fully represent the initial data represents an advantage in respect with previous studies. Chastrette *et al.*[15] used eight molecular descriptors for a database of 83 solvents. The highly correlated descriptors could be represented by three principal components with an information loss of 18%. Alan R. Katritzky *et al.*[49] further expanded this approach, initially considering 40 descriptors for 40 solvents, and later 100 descriptors for 774 solvents,[16] all descriptors calculated with QSPR. Resulting in a representation of the solvents with three principal components accounting for over 60% of the variation. Stairs *et al.*[50] applied PCA on solvent spectra and equilibrium rates of reaction. The methodology advanced to a stage where refinement was necessary, with the main challenge being the sparsity of data in certain regions of the solvent spectrum. Diorazio *et al.*[8] addressed this issue *via* the implementation of an interactive tool reducing the solvent representation from 17 descriptors to six PCs while capturing 87.9% of the variation. Durand *et al.*[17] followed the same approach on the σ-potential. However, the decorrelation process in PCA is not sufficiently robust to fully decompose curves.[50] The need for 5 PCs to capture 98.6% of the variation (Fig. 3) limits its usefulness for solvent visualization, clustering and the interpretation of the resulting PCs. The effects on the clusters proposed by Durand *et al.*[17] and later reworked by Moity *et al.*[18] are presented in Fig. 7. The clusters appear correctly positioned within the FPCA mapping: apolar solvents reside in the top-left corner, while aprotic and pair-donor bases are situated at the top, shifting leftward as polarity increases. Solvents characterized by dual positive and negative surface charge screening (amphiprotic) are correctly located in the center.

Open Access Article. Published on 19 February 2026. Downloaded on 6/19/2026 1:24:24 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

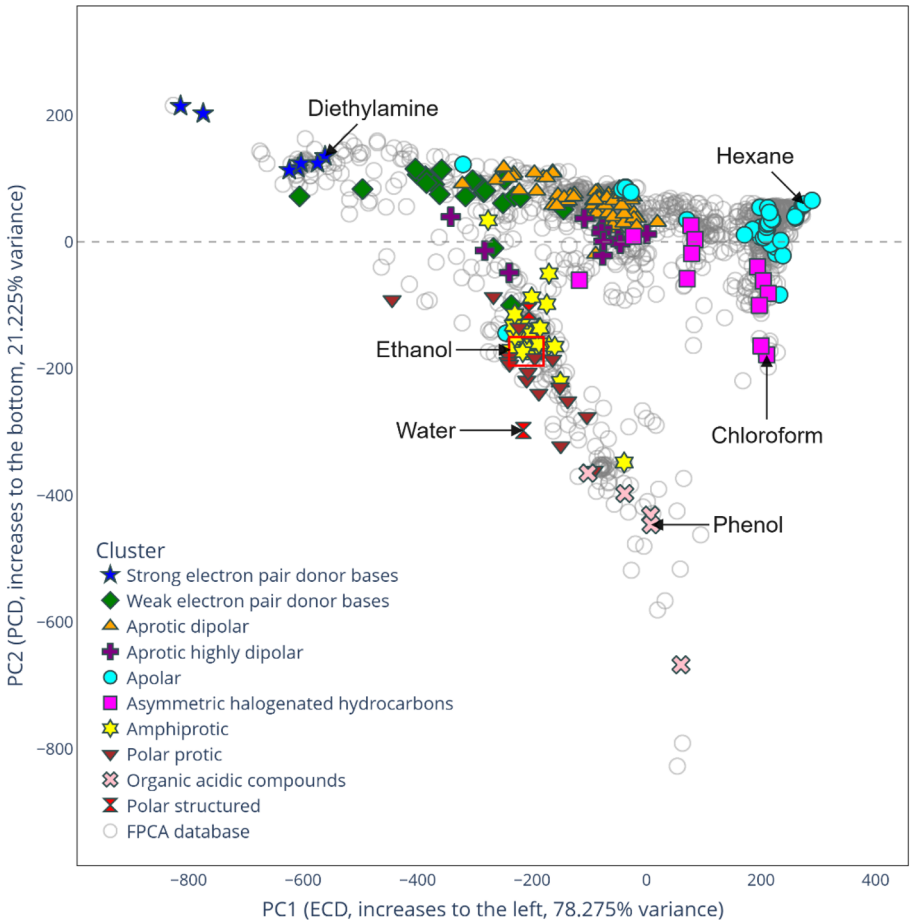

Fig. 7 Solvents of our database in the FPCA space. Solvents have been separated by the clusters assigned by Moity *et al.* for comparison. The area within the red rectangle is visualized in Fig. 7.

However, the amphiprotic and polar protic clusters show significant overlap, indicating that they share the same σ-potential shape and that distinguishing these zones into two separate families may be unnecessary. Asymmetric halogenated hydrocarbons span a wide region, as their σ-potentials vary from acidic to basic, a nuance that is visualized more effectively in the FPCA mapping than in previous models.

While there is also overlap among aprotic families, the displacement to the left in the FPCA map clearly distinguishes molecules with higher ECD more effectively than standard PCA. These limitations in the original classification stem from Durand *et al.*'s use of the third principal component to represent lipophilicity, which inadvertently results in a loss of differentiation regarding polarization.

Furthermore, applying clustering to the results of standard PCA leads to the misclassification of several solvents that exhibit behavior distinct from their assigned clusters. For example, decamethylcyclopentasiloxane, r,r-diisopropyleneglycol, and a series of esters near the aprotic dipolar family are incorrectly classified as apolar. Triethyleneglycol is misidentified as a weak electron-pair donor base, while aniline is grouped with asymmetric halogenated hydrocarbons. Additionally, two molecules from the amphiprotic family are superimposed onto other families: tetraethylene glycol is found near the electron-pair donor bases, and oleic acid is positioned as an organic acidic compound.

The positioning of these solvents within the FPCA mapping shows an improvement on the representation of their polarization profile and highlight the ability of FPCA as a better tool for dimensionality reduction compared to PCA besides requiring only two dimensions to capture the information contained in the σ-potential.

Furthermore, the use of a more extensive database compared to Moity *et al.* (solvents in gray had not been used in Durand's study) illustrates the lack of a defined families. Instead, a continuum is observed and the description of regions that gradually change their polarization behavior.

The effects of lipophilicity, reflected by the amount of apolar segments in the σ-profile/surfaces, can be observed in the FPCA plane (Fig. 8) with the alcohols and the effect of the alkyl length on their position in the FPCA plane. As the alkyl chain length increases in higher alcohols, the polarization of the O–H bond, and consequently the partial charges (δ^-) on oxygen and (δ^+) on hydrogens gradually decrease resulting in lowering acidity. This trend is attributed to the increasing electron-donating inductive effect (+I) of the increasingly longer alkyl chain in the molecule. It is worth noting that phenol does not appear in the region corresponding to

© 2026 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 19 February 2026. Downloaded on 6/19/2026 1:24:24 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

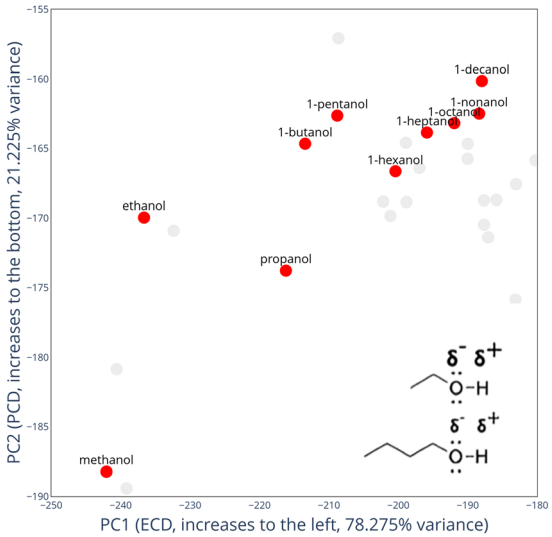

Fig. 8 Visualization of alcohols in the FPCA space and the effect of the inductive effect of the alkyl chain in partial charge.

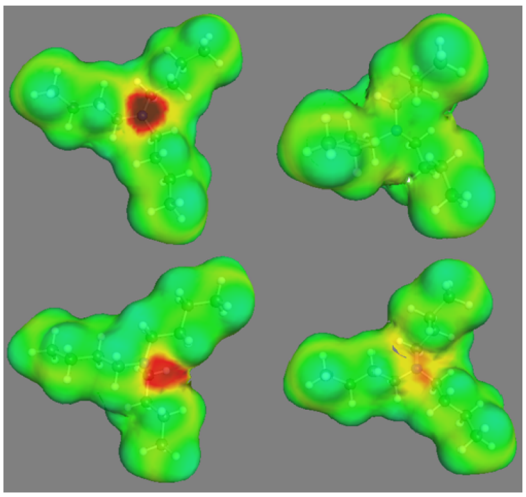

Fig. 9 σ-Surface of 4 different tributylamine's conformers.

alcohols (Fig. 8) as its higher acidity results from the delocalization of the oxygen electron pair into the phenyl cycle through an electron-donating mesomeric. This better electronic distribution throughout the molecule also explained the decrease in ECD compared to acyclic alcohols.

The treatment of the data by PCA as vectors of discrete observations, results in considering the σ-potential as a series of individual points, without accounting for the correlation along the whole curve. This can explain the observed information loss. FPCA avoids this issue by considering the σ-potential not as a set of discrete points but as a single functional object. This approach is, however, limited by how the σ-potential is constructed. Since the geometric information of the molecule is lost when the σ-surface is transformed into the σ-profile and σ-potential, steric effects are not fully captured. Molecules lacking accessible polar regions can display either the expected behavior or the opposite. For example, the tertiary amine tributylamine is known to behave as a weaker base than its primary and secondary homologues due to the steric hindrance imposed by its three butyl chains. Nonetheless, as these chains do not completely shield the region above the nitrogen atom, the molecule is represented as having a high electron charge density (ECD), although it is not readily accessible to capture a proton (Fig. 9).

Since activity coefficients can be derived directly from the σ-potential, solvents with similar σ-potentials may serve as effective substitutes. Accordingly, the distance between points in the FPCA space provides a valuable methodology to identify alternatives, as solvents located closer together are more likely to share similar σ-potentials and solvent–solute affinities. This methodology represents a preliminary tool for identifying solvent substitutes. It also serves as an alternative to COSMO-RS relative solubility, particularly when solutes are too complex for the calculation of required COSMO surface input, a limitation that is not present in our methodology.

Case studies

To assess the validity of the proposed mapping and its methodology for identifying alternative solvents to replace toxic ones, several case studies will be analyzed in the following, including nitrocellulose, ester gum, and rapeseed oil, selected owing to the availability of solubilization data.

Case study 1: nitrocellulose

Nitrocellulose (Fig. 10) is a compound widely used in the production of lacquers, explosives and celluloids. Due to its importance in coatings, inks and propellants, the global nitrocellulose market is projected to reach 1.39 billion USD by 2030. The production of nitrocellulose involves the nitration of cellulose, typically derived from wood pulp or cotton fiber due to its natural high purity cellulose content.[51] The nitration process introduces nitro groups (–NO_2) onto the cellulose backbone, significantly altering its chemical properties and solubility behavior. Chastrette *et al.*[15] and Durand *et al.*[17] validated their approaches using the search of green solvents for nitrocellulose solubilization as benchmark case. Hansen's solubility data[47] for nitrocellulose (type H23, A. Hagedorn & Co) in 80 solvents (complete list in SI) represented in the FPCA plane (Fig. 10). For the experiments, 0.5 g of polymer was placed in a test tube with 5 mL of solvent. Hansen's qualitative solubility scoring was carried out by considering two categories: high (scores 1–2) and low (scores 3–6) solvent–polymer affinities.

Experimental results revealed that solvents capable of dissolving nitrocellulose grouped in the same regions as the clusters described by Durand *et al.*[17] The most effective solvents are aprotic ranging from fairly to quite polar located in the upper section of the FPCA space like DMSO, THF and propylene carbonate (Fig. 11). In contrast, those in the apolar region have no effect on the solubilization of nitrocellulose. This behavior was attributed to their apolar surface area, which cannot effectively interact with the charged nitro groups present in nitrocellulose. Some protic solvents in the center of the FPCA plane exhibited good solubility. Further analysis of σ-profiles

Open Access Article. Published on 19 February 2026. Downloaded on 6/19/2026 1:24:24 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

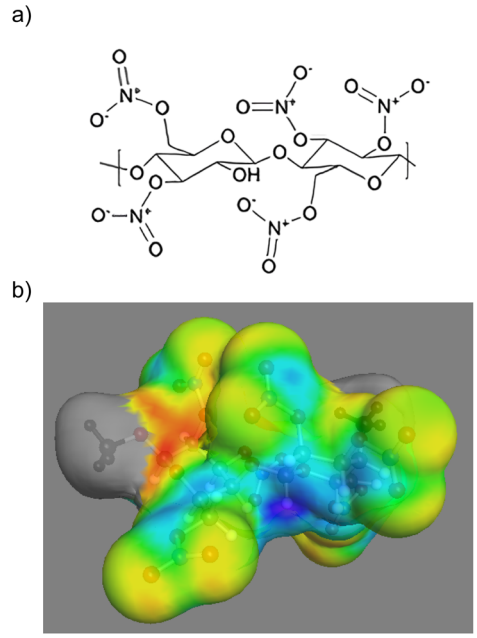

Fig. 10 (a) Chemical structure and (b) σ-surface of nitrocellulose.

revealed that polar solvents with smaller apolar surface area like methanol, 2-methoxyethanol, and diisopropylene glycol, interact more effectively with the polar nitro groups ($-NO_2$) and hydroxyl sites present in nitrocellulose. On the other hand, solvents with poor solubility in the aprotic polar region are cases where although there are polarized regions those regions are in symmetrical positions cancelling their dipolar moments, as in the case of dioxane and diethyl sulfide. Aniline is another interesting case, while it presents electron-donor and electron accepting regions, aniline tends to form molecular aggregation in liquid state due to hydrogen bonds,[52] reducing its availability to solvate nitrocellulose chains.

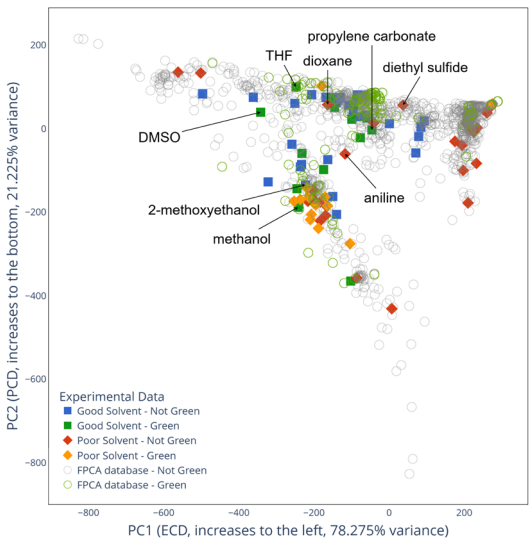

Fig. 11 FPCA with 80 solvents tested experimentally for the solubilization of nitrocellulose.

Case study 2: ester gum

Ester gum is a pale colored thermoplastic resin used in several industrial applications such as adhesives, inks, foods, beverages, coatings and personal care products. Its solubility in organic solvents is critical for its role as plasticizer,[53] enhancing film formation and durability. It is a derivative of rosin, composed primarily of resin acids such as abietic acid (Fig. 12), palustric acid and neoabietic acid. It can be dissolved in solvents like hexane, acetone and alcohols, making it suitable for lacquer formulations.

Hansen's solubility data[47] of Ester Gum BL (Hercules Incorporated) in various solvents (complete list in SI) was determined, with 80 of these solvents included in the present classification (Fig. 13). For the experiments, 0.5 g of polymer was placed in a laboratory glass tube with 5 mL of solvent. Hansen's qualitative solubility scoring was simplified into two categories: high (scores 1–2) and low (scores 3–6) solvent–polymer affinities.

Ester Gum is mostly apolar despite the presence of one or two –OH groups which allows it to be soluble in polar and aprotic solvents (diethyl ether, pyridine). Some protic solvents (2-octanol, 2-ethyl-1-butanol) can also dissolve ester gum, provided they have a sufficient proportion of non-polar surface area. For example, short-chain alcohols (such as methanol, ethanol, tert-butanol, and butanol, *etc.*) are unable to dissolve this molecule. In this case, the diversity of solvent regions covered by ester gum on the FPCA map (Fig. 13) reflects its versatility, which opens the possibility of applying additional filters (less ecotoxicity, bio-based content, specific interaction with other ingredients, *etc.*) to narrow down the list of candidate solvents depending on the target application.

Case study 3: rapeseed oil

Rapeseed oil is one of the most consumed oils in the world. It is extensively used for human consumption, and as a replacement

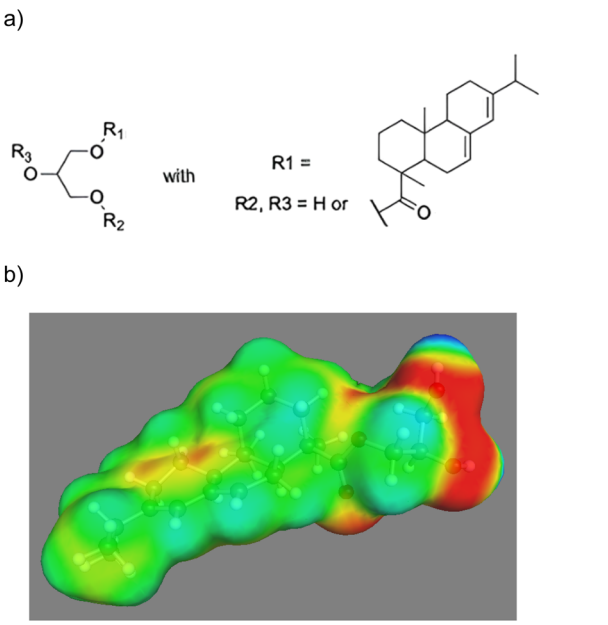

Fig. 12 (a) Chemical structure of molecules composing Ester Gum. (b) σ-surface of abietic acid monoglycerol.

© 2026 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 19 February 2026. Downloaded on 6/19/2026 1:24:24 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

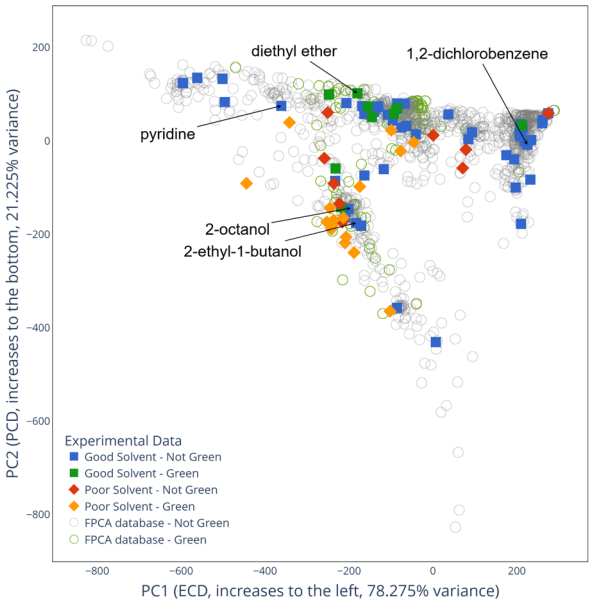

Fig. 13 FPCA space for 80 solvents tested experimentally for the solubilization of Ester Gum.

for petroleum-based oil products. It is mainly extracted using hexane, which poses several hazards to human health and the environment.[54] Rapeseed oil is composed of triglycerides derived mostly from oleic acid. The chemical structure and the σ-surface of triolein are displayed in Fig. 14. The Hansen solubility parameters of rapeseed oil and the solubility sphere were experimentally determined for 56 solvents (complete list in SI) from solutions having 25% weight of rapeseed oil. A score of 1 was given for homogenous monophasic mixtures, whereas biphasic, highly turbid, or opaque systems indicating partial or no solubilization were assigned a score of 0.

As shown in Fig. 14, triolein, which accounts for around 70% of the total fat content of rapeseed oil, has large low-polarized areas due to its three long alkenyl chains. Thus, aprotic apolar like cyclohexane indicated on the FPCA map of Fig. 15, are good candidates for dissolving the oil due to interaction with the long carbon chains in triolein. The presence of three ester groups attracts electron density through mesomeric effect and generating electron-depleted zones around neighboring hydrogens. These zones act as anchoring points for aprotic polar solvents with mild to quite high polarity. Consequently, molecules located in the upper region of the PCA space (Fig. 15), with a sufficiently large apolar surface area to interact with carbon chains represent promising solvent candidates, such as cyclohexanone, dioxolane, and 2-methyltetrahydrofuran.

Greener solvents located near regions densely populated by "good toxic solvents" may offer promising alternatives in industrial applications to comply with international regulations, such as REACH, thereby reducing adverse effects on human health and the environment. Greener solvents, which often include bio-based, recyclable, eco-friendly and/or low-toxicity compounds, are increasingly incorporated into manufacturing processes, including pharmaceuticals and

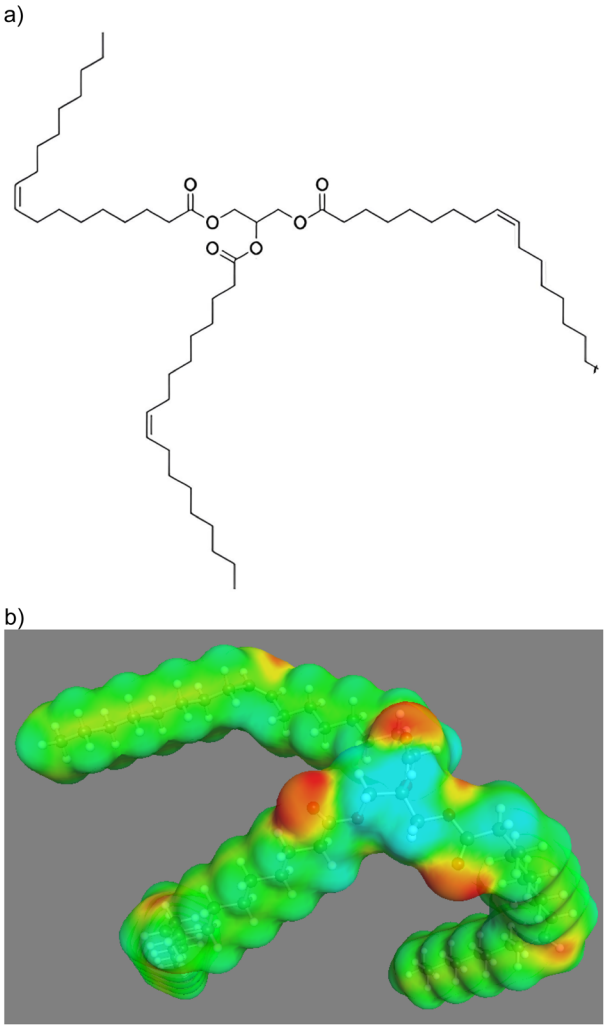

Fig. 14 (a) Chemical structure and (b) σ-surface of triolein.

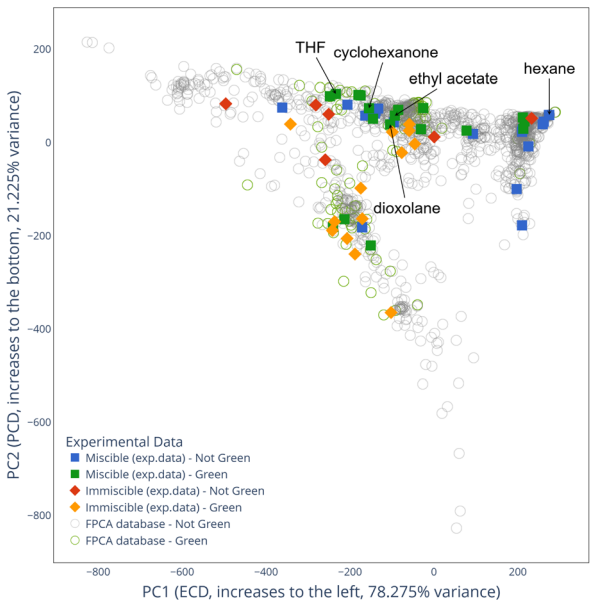

Fig. 15 FPCA space for 56 solvents tested experimentally for the solubilization of rapeseed oil.

Open Access Article. Published on 19 February 2026. Downloaded on 6/19/2026 1:24:24 PM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

agrochemicals for coatings processes and polymers production. Their adoption not only supports sustainability goals but also enhances process safety, reduces waste and improves occupational health conditions. Alternative solvents can be found using this cartography and then selected according to the properties required for the intended application other than those of solubilization.

Conclusions

In the context of searching for alternative solvents, we propose a new and simplified mapping approach that can be used for the substitution of toxic organic solvents. This work highlights the potential of Functional Principal Component Analysis (FPCA) as an advanced method to represent and interpret solvent solubility capacity using the σ-potential molecular descriptor computed from COSMO-RS. Unlike classical Principal Component Analysis (PCA), which treats molecular descriptors as discrete and uncorrelated variables, FPCA exploits the functional nature of the σ-potential to extract chemically interpretable and physically meaningful dimensions, particularly charge density ECD and PCD while preserving the inherent structural continuity of molecular profiles. The first FPCA component primarily corresponds to the solvent's ability to donate electrons (ECD) and, to some extent, to accept hydrogen bonds. The second component corresponds to the solvent's positive charge density (PCD) and reflects the solvent's acidity. The resulting two-dimensional FPCA space offers chemically intuitive representation of solvent behaviour in a two-dimensional space, allowing the identification of distinct regions: apolar solvents clustered in the upper right quadrant, aprotic solvents with increasing ECD towards the left, and protic solvents exhibiting gradient trends along the vertical axis.

Comparisons with earlier approaches that used PCA on multiple discrete descriptors confirm the superiority of FPCA in both interpretability and data fidelity capturing 99.5% of the variance with only two components. By contrast, PCA applied to the same σ-potential dataset required four components to retain ~96% variance and relied on a low-variance dimension as a major component for solvent differentiation, which led to overlapping clusters.

The utility of the FPCA-derived solvent space was validated through three case studies: nitrocellulose, ester gum, and rapeseed oil. These examples demonstrate FPCA's predictive capacity for solvent efficacy, as the relative positioning of good and poor solvents consistently confirmed the explanatory power of the FPCA dimensions. For instance, solvents effective in dissolving nitrocellulose clustered in the zone of moderately high ECD and mild acidity, consistent with its polar nature. By contrast, the dissolution behaviour of ester gum and rapeseed oil, both more apolar in nature, was predominantly governed by solvent lipophilicity and auxiliary polarity features, as reflected by their distinct spatial distribution within the FPCA framework.

Because the FPCA space preserves the functional shape and nature of the σ-potential, solvents located in close proximity can be expected to behave similarly, even when not previously tested experimentally. This opens promising opportunities for identifying greener and safer alternatives to traditional solvents using distance metrics within this reduced two-dimensional space. The clustering of known "green" solvents in proximity to functional equivalents suggests that the FPCA space can serve as a preselection tool for candidates that align with regulatory (*e.g.*, REACH, CHEM21) and sustainability goals. The addition of new solvents, either as pure compounds or as multicomponent mixtures, can be easily carried out once their associated σ-potentials have been obtained using COSMO-RS. These solvents can then be added to the database, and FPCA can be reapplied to obtain their corresponding coordinates in the 2D-mapping.

The direct relationship between activity coefficients and σ-potentials confirms that solvents with similar profiles may serve as substitutes. Our results demonstrate that the distance in FPCA space successfully captures this similarity, providing a method to identify alternatives based on proximity. This framework represents a novel preliminary tool for selecting alternative solvents. Crucially, it provides a solution for systems where COSMO-RS relative solubility cannot be applied due to the complexity of generating solute COSMO surfaces, as our methodology does not require these inputs.

Nonetheless, this approach has limitations. Although FPCA captures the shape of the chemical potential across the solvent surface, it does not explicitly encode molecular geometry which can displace molecules from their expected behaviour as in the case of tributylamine. The mapping obtained from FPCA also does not account for intermolecular interactions leading to the formation of dimers or aggregates. Moreover, the transformation from σ-surface to σ-profile and σ-potential entails information loss, omitting specific three-dimensional conformational and polarization details that may influence solvent behaviour in complex systems. In such cases, the respective contributions of conformer selection and specific hydrogen-bond coordination must be evaluated to ensure an optimal application of the methodology. These limitations highlight the need for future extensions that incorporate 3D geometric data to capture combinatorial contributions or explicit consideration of the σ-surface and σ-profile, thereby transforming the plane into a three-dimensional representation.

The methodology also needs to be validated in predictive modelling in QSPR models or ML approaches to transition from a qualitative description to a quantitative predictive capacity. Finally, σ-moments descriptors that have got renewed interest[38] would be a suitable benchmark to compare our FPCA approach although the σ-moments relate to σ-potential and σ-profiles while our PC1 and PC2 relate only to σ-potential.

Author contributions

Luis E. Ramirez Cardenas: conceptualization, investigation, methodology, formal analysis, visualization, writing (original draft & editing). Rachid Ouaret: methodology, writing (review & editing). Vincent Gerbaud: writing (review & editing). Ivonne Rodriguez Donis: project administration, resources, supervision, writing (review & editing). Sophie Thiebaud-Roux: project administration, resources, supervision, writing (review & editing).

© 2026 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 19 February 2026. Downloaded on 6/19/2026 1:24:24 PM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Conflicts of interest

There are no conflicts to declare.

Data availability

Data supporting this study are included within supplementary information (SI). Supplementary information: database – list of molecules and their σ-potentials used in the FPCA reduction. The clusters assigned by Moity *et al.* are included, nitrocellulose and ester gum – list of tested solvents by Hansen *et al.* and their corresponding scores ranging from 1 (best) to 6 (worst); triolein – list of tested solvents for the solubilization of triolein, a score of 1 was assigned for full miscibility and 0 otherwise; green solvents – list of solvents classified as green by James Clark, Moity *et al.*, and the developed list from our laboratory. See DOI: https://doi.org/10.1039/d5ra08246c.

Acknowledgements

This work has benefited from state aid managed by the French National Research Agency (ANR) under the project ANR-21-CE05-0031-01 as well as the "Investissements d'Avenir" program with the reference ANR-18-EURE-0021.

Notes and references

1 N. Winterton, *Clean Technol. Environ. Policy*, 2021, **23**, 2499–2522.
2 V. O. C. Solvents Emissions Inventories Technical Paper 2021, https://www.esig.org/post/solvents-voc-emission-inventories-technical-paper-2021/, (accessed May 5, 2025).
3 Understanding REACH - ECHA, https://echa.europa.eu/regulations/reach/understanding-reach, (accessed May 5, 2025).
4 D. Prat, J. Hayler and A. Wells, *Green Chem.*, 2014, **16**, 4546–4551.
5 C. Jiménez-González, A. D. Curzons, D. J. C. Constable and V. L. Cunningham, *Clean Technol. Environ. Policy*, 2004, **7**, 42–50.
6 A. D. Curzons, D. C. Constable and V. L. Cunningham, *Clean Prod. Process.*, 1999, **1**, 82–90.
7 R. K. Henderson, C. Jiménez-González, D. J. C. Constable, S. R. Alston, G. G. A. Inglis, G. Fisher, J. Sherwood, S. P. Binks and A. D. Curzons, *Green Chem.*, 2011, **13**, 854–862.
8 L. J. Diorazio, D. R. J. Hose and N. K. Adlington, *Org. Process Res. Dev.*, 2016, **20**, 760–773.
9 D. Prat, O. Pardigon, H.-W. Flemming, S. Letestu, V. Ducandas, P. Isnard, E. Guntrum, T. Senac, S. Ruisseau, P. Cruciani and P. Hosek, *Org. Process Res. Dev.*, 2013, **17**, 1517–1525.
10 Solvent Tool – ACSGCIPR, https://acsgcipr.org/tools/solvent-tool/, (accessed May 5, 2025).
11 C. S. Slater and M. Savelski, *J. Environ. Sci. Health Part A*, 2007, **42**, 1595–1605.
12 C. Capello, S. Hellweg and K. Hungerbühler, *J. Ind. Ecol.*, 2008, **12**, 111–127.
13 F. Pin, J. Picard and S. Dhulut, *Org. Process Res. Dev.*, 2025, **29**, 1715–1726.
14 F. P. Byrne, S. Jin, G. Paggiola, T. H. M. Petchey, J. H. Clark, T. J. Farmer, A. J. Hunt, C. Robert McElroy and J. Sherwood, *Sustainable Chem. Processes*, 2016, **4**, 7.
15 M. Chastrette, M. Rajzmann, M. Chanon and K. F. Purcell, *J. Am. Chem. Soc.*, 1985, **107**, 1–11.
16 A. R. Katritzky, I. Tulp, D. C. Fara, A. Lauria, U. Maran and W. E. Acree, *J. Chem. Inf. Model.*, 2005, **45**, 913–923.
17 M. Durand, V. Molinier, W. Kunz and J.-M. Aubry, *Chem. Eur J.*, 2011, **17**, 5155–5164.
18 L. Moity, M. Durand, A. Benazzouz, C. Pierlot, V. Molinier and J.-M. Aubry, *Green Chem.*, 2012, **14**, 1132–1145.
19 J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*, Springer, New York, NY, 2005.
20 J. A. Rice and B. W. Silverman, *J. Roy. Stat. Soc. B*, 1991, **53**, 233–243.
21 A. Klamt, *J. Phys. Chem.*, 1995, **99**, 2224–2235.
22 J. H. Clark, A. Hunt, C. Topi, G. Paggiola and J. Sherwood, *Sustainable Solvents: Perspectives from Research, Business and International Policy*, Green Chemistry Series, Royal Society of Chemistry, 2017.
23 C. Reichardt and T. Welton, *Solvents and Solvent Effects in Organic Chemistry*, John Wiley & Sons, 2011.
24 J. H. Hildebrand, *Chem. Rev.*, 1949, **44**, 37–45.
25 C. M. Hansen, *The Three Dimensional Solubility Parameter and Solvent Diffusion Coefficient: Their Importance in Surface Coating Formulation*, Danish Technical Press, 1967.
26 M. J. Kamlet, R. M. Doherty, J.-L. M. Abboud, M. H. Abraham and R. W. Taft, *J. Pharm. Sci.*, 1986, **75**, 338–349.
27 S. Jin, F. Byrne, C. R. McElroy, J. Sherwood, J. H. Clark and A. J. Hunt, *Faraday Discuss.*, 2017, **202**, 157–173.
28 J. L. Franklin, *Ind. Eng. Chem.*, 1949, **41**, 1070–1076.
29 A. S. Hukkerikar, B. Sarup, A. Ten Kate, J. Abildskov, G. Sin and R. Gani, *Fluid Phase Equilib.*, 2012, **321**, 25–43.
30 S. Linke, K. McBride and K. Sundmacher, *ACS Sustainable Chem. Eng.*, 2020, **8**, 10795–10811.
31 E. Benfenati, A. Manganaro and G. C. Gini, *Pai@ Ai* IA*, 2013, **1107**, 21–28.
32 A. Klamt and G. Schüürmann, *J. Chem. Soc., Perkin Trans.*, 1993, **2**, 799–805.
33 R. Höfer, *Sustainable Solutions for Modern Economies*, Royal Society of Chemistry, Cambridge, UK, 2009, pp. 407–423.
34 P. M. Piccione, J. Baumeister, T. Salvesen, C. Grosjean, Y. Flores, E. Groelly, V. Murudi, A. Shyadligeri, O. Lobanova and C. Lothschütz, *Org. Process Res. Dev.*, 2019, **23**, 998–1016.
35 A. Klamt and F. Eckert, Rational approaches to drug design, *Prous Science SA*, 2001, 195–205.
36 C. Mehler, A. Klamt and W. Peukert, *AIChE J.*, 2002, **48**, 1093–1099.
37 COSMOtherm, https://www.3ds.com/products/biovia/cosmo-rs/cosmotherm, (accessed January 19, 2026).
38 F. Y. M. Salih, D. O. Abranches, E. J. Maginn and Y. J. Colón, *Digital Discovery*, 2025, **4**, 2711–2723.

Open Access Article. Published on 19 February 2026. Downloaded on 6/19/2026 1:24:24 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

39 BIOVIA, https://www.3ds.com/fr/products/biovia, (accessed September 4, 2025).
40 C. C. Pye and T. Ziegler, *Theor. Chem. Acc.*, 1999, **101**, 396–408.
41 S.-T. Lin and S. I. Sandler, *Ind. Eng. Chem. Res.*, 2002, **41**, 899–913.
42 T. Gerlach, S. Müller, A. G. de Castilla and I. Smirnova, *Fluid Phase Equilib.*, 2022, **560**, 113472.
43 P. Gramatica, *QSAR Comb. Sci.*, 2006, **25**, 327–332.
44 TURBOMOLE | Program Package for Electronic Structure Calculations, https://www.turbomole.org/, (accessed June 13, 2025).
45 A. Schäfer, A. Klamt, D. Sattel, J. C. W. Lohrenz and F. Eckert, *Phys. Chem. Chem. Phys.*, 2000, **2**, 2187–2193.
46 C. Ramos-Carreño, J. L. Torrecilla, M. Carbajo-Berrocal, P. Marcos and A. Suárez, *J. Stat. Software*, 2024, **109**, 1–37.
47 C. M. Hansen, *Hansen Solubility Parameters: A User's Handbook, Second Edition*, CRC Press, Boca Raton, 2nd edn, 2007.
48 M. Nehmeh, I. Rodriguez-Donis, V. Gerbaud and S. Thiebaud-Roux in *Computer Aided Chemical Engineering*, ed. A. C. Kokossis, M. C. Georgiadis and E. Pistikopoulos, Elsevier, 2023, vol. 52, pp. 1939–1944.
49 A. R. Katritzky, T. Tamm, Y. Wang and M. Karelson, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 692–698.
50 R. A. Stairs and E. Buncel, *Can. J. Chem.*, 2006, **84**, 1580–1591.
51 E. Morris, C. R. Pulham and C. A. Morrison, *RSC Adv.*, 2023, **13**, 32321–32333.
52 A. Jumabaev, B. Khudaykulov, I. Doroshenko, H. Hushvaktov and A. Absanov, *Vib. Spectrosc.*, 2022, **122**, 103422.
53 B. K. Brown, The Use of Plasticizers in Lacquers, https://pubs.acs.org/doi/pdf/10.1021/ie50186a005, (accessed June 17, 2025).
54 J. Shen, Y. Liu, X. Wang, J. Bai, L. Lin, F. Luo and H. Zhong, *Nutrients*, 2023, **15**, 999.

© 2026 The Author(s). Published by the Royal Society of Chemistry