Computational and Structural Biotechnology Journal 21 (2023) 3234–3247

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Review article

An overview of descriptors to capture protein properties – Tools and perspectives in the context of QSAR modeling

J. Emonts [a], J.F. Buyel [b,c,*,1]

[a] Fraunhofer Institute for Molecular Biology and Applied Ecology IME, Germany
[b] University of Natural Resources and Life Sciences, Vienna (BOKU), Department of Biotechnology (DBT), Institute of Bioprocess Science and Engineering (IBSE), Muthgasse 18, 1190 Vienna, Austria
[c] Institute for Molecular Biotechnology, Worringerweg 1, RWTH Aachen University, 52074 Aachen, Germany

ARTICLE INFO

Article history:
Received 13 March 2023
Received in revised form 23 May 2023
Accepted 23 May 2023
Available online 24 May 2023

Keywords:
Prediction of molecular features
Protein structure complexity
Quantitative structure activity relationship
Scalar parameters
Shape and surface properties

ABSTRACT

Proteins are important ingredients in food and feed, they are the active components of many pharmaceutical products, and they are necessary, in the form of enzymes, for the success of many technical processes. However, production can be challenging, especially when using heterologous host cells such as bacteria to express and assemble recombinant mammalian proteins. The manufacturability of proteins can be hindered by low solubility, a tendency to aggregate, or inefficient purification. Tools such as *in silico* protein engineering and models that predict separation criteria can overcome these issues but usually require the complex shape and surface properties of proteins to be represented by a small number of quantitative numeric values known as descriptors, as similarly used to capture the features of small molecules. Here, we review the current status of protein descriptors, especially for application in quantitative structure activity relationship (QSAR) models. First, we describe the complexity of proteins and the properties that descriptors must accommodate. Then we introduce descriptors of shape and surface properties that quantify the global and local features of proteins. Finally, we highlight the current limitations of protein descriptors and propose strategies for the derivation of novel protein descriptors that are more informative.

Contents

Abbreviations: CDR, complementarity determining regions; IDR, intrinsically disordered regions; IDP, PDB, intrinsically disordered proteinsProtein Data Bank; QSAR, quantitative structure–activity relationship; RuBisCO, ribulose-1,5-bisphosphate carboxylase-oxygenase; SASA, (solvent) accessible surface area; vdW, van der Waals

* Corresponding author at: University of Natural Resources and Life Sciences, Vienna (BOKU), Department of Biotechnology (DBT), Institute of Bioprocess Science and Engineering (IBSE), Muthgasse 18, 1190 Vienna, Austria.
E-mail address: johannes.buyel@rwth-aachen.de (J.F. Buyel).
[1] ORCID: 0000-0003-2361-143X

https://doi.org/10.1016/j.csbj.2023.05.022

1. Introduction

Proteins are important components of food and feed, providing amino acids that facilitate growth and maintain health [1,2]. Many active pharmaceutical ingredients (APIs) are also proteins, including subunit vaccines, antibodies, blood products and replacement enzymes [3–6]. Furthermore, numerous industrial processes rely on enzymes for the catalysis of chemical reactions that are too complex and/or expensive for total chemical synthesis, such as the production of paclitaxel [7,8]. Unfortunately, the manufacturability of proteins can be hindered by low solubility, a tendency to aggregate or inefficient separation from other proteins, nucleic acids or small molecules [9,10]. Several *in silico* approaches have therefore been developed to address these challenges, including rational protein engineering and model-based bioprocess development [11–13].

The approaches typically require the quantification of properties such as protein surface charge distribution to facilitate *in silico* screening and optimization. The quantitation can be achieved using descriptors that are discrete scalars (e.g., the number of positive surface charges), continuous scalars (e.g., the isoelectric point, pI), or vectors that capture molecular features and thus act as surrogates for actual properties [14], i.e. they reduce the dimensionality of an object, here: a protein. For example, instead of using the explicit charge of each surface-expose functional group of a protein at a given pH (e.g. 19 negative and 3 positive charges at pH 8.5) one can use the isoelectric point or net charge (e.g. 16 negative at pH 8.5) as global charge descriptors.

A key limitation of this dimensionality reduction is that, depending on their definition, descriptors may fail to capture protein complexity and the properties relevant for a specific application, e.g. the prediction of chromatographic protein separation. Specifically, the use of descriptors is characterized by an information loss. In particular, the complex interactions between different protein properties, such as shape and surface charge, can be challenging to capture in a descriptor. Hence, it is advisable to choose or develop descriptors using domain knowledge about the underlying molecular mechanisms of a specific application to ensure that the relevant information is maintained after dimensionality reduction.

The descriptor concept is well established in chemometrics, where it is used to characterize small-molecule compounds with a mass of less than 5000 kDa [15–17]. For example, typical small-molecule descriptors include the number of nitrogen atoms, the number of phenol rings, and the number of hydroxyl groups. Several suggestions for adapting the concept to proteins have been made [18–20] and discussing all of them in detail is beyond the scope of this review.

One main application of descriptors in case of proteins is quantitative structure–activity relationship (QSAR) modeling [21,22]. QSAR models correlate molecular properties (e.g., descriptors such as the isoelectric point) with experimental data reporting measurable properties such as solubility, ligand interaction, or oligomerization [23–25]. If descriptors can be calculated ab initio, for example based on the protein's shape and surface properties, such correlations can facilitate a priori predictions about molecular properties that have not been determined experimentally. Accordingly, manufacturability can be assessed during early product design and bioprocess development so that resources can be allocated to the most promising candidates.

Whereas descriptors provide meaningful information about small molecules, they quickly become ambiguous and thus irrelevant as the molecular mass and complexity increase. The ambiguity is more pronounced in proteins because they are polymers composed of a limited number of different but repeating monomeric building blocks. For example, two proteins like endo-beta-1,4-glucanase B (UniProt ID O74706) and glyceraldehyde-3-phosphate dehydrogenase A (UniProt ID P0A9B2) may have a similar atomic composition (3112 vs 3126 carbon atoms) and molecular mass (69.9391 kDa vs 70.8043 kDa) but adopt completely different structures because of differences in the primary sequence, whereas two proteins with different compositions but a limited degree of sequence homology may form similar structures [26–28]. Accordingly, descriptors must be tailored for proteins and the specific process or application problem. In the context of chromatography, the problem may be the determination of protein–ligand interactions that can in turn be used to predict isotherm parameters and, ultimately, separation conditions [29]. Protein descriptors can be discriminated into various types depending on the application context. Here, we classify protein descriptors into two main groups: surface descriptors and those related to shape, size or structure (for brevity, we describe these as surface descriptors and shape descriptors hereafter). These categories are useful in the context of modeling protein separation, e.g. during chromatography, as, for example, shape will determine diffusion whereas surface properties will determine sorption processes. However, other classifications may be more suitable in a different context and some descriptors are ambiguous because they could be assigned to more than one group. For example, the isoelectric point captures aspects of the protein surface charge and the (solvent) accessible surface area ((S)ASA), which is related to shape.

In this review, we describe the complex features of proteins that must be captured by descriptors and the introduce the various families of shape and surface descriptors that are used, highlighting their applications and limitations, specifically in the context of QSAR models. Finally, we discuss properties that are underrepresented by current descriptors and opportunities to cover them when creating novel descriptors. We focus our discussion on bioprocess development, specifically downstream processing, such as chromatographic purification. We will not cover membrane proteins or the methods of descriptor selection because this must be carried out in the context of specific QSAR applications [30]. Also, descriptors covering biological function will not be discussed.

2. Determinants of protein structure complexity that have to be covered by descriptors

The purpose of protein descriptors is to capture chemical information and relevant properties of the three-dimensional structure in a simple, often one-dimensional form, for example as a scalar or set of scalars like the *k* largest positively charged surface patches. Therefore, in an abstract sense, calculating the numeric value of a descriptor is a complexity reduction task. The challenge is to minimize the information loss when extracting the important aspects of

protein properties, for example the surface charge (distribution), while reducing the dimensionality of the original object [31].

To illustrate the challenge of complexity reduction, we will briefly describe the important structural elements of proteins. The four well-known structural levels are the primary structure (amino acid sequence), the secondary structure (repeating local configurations such as α-helices, β-sheets and β-turns, generally held together by hydrogen bonds) or lack thereof (random coil), the tertiary structure (overall configuration or "fold", defined by the path of the polypeptide backbone through space, which is stabilized by disulfide and hydrogen bonds as well as ionic and hydrophobic interactions) and the quaternary structure (assembly of polypeptide subunits into a higher-order structure) [32]. The quaternary structure is also described as oligomerization and may comprise multiple copies of the same polypeptide (homomultimer) or different polypeptides (heteromultimer) [33]. Various degrees of oligomerization exist, starting with the monomer (e.g., bovine serum albumin; UniProt ID P02769) and dimer (e.g., human alcohol dehydrogenase class 3; UniProt ID P11766), but building to complex assemblies such as hetero-16-mers (e.g., ribulose-1,5-bisphosphate carboxylase-oxygenase (RuBisCO); UniProt IDs P00876 and P69249) and hetero-24-mers (e.g., human ferritin; UniProt IDs P02792 and P02794). Importantly, oligomerization can change the properties of the monomeric unit compared to the free monomer. The oligomer always has a larger mass and size, but the monomeric shape and surface charge may also change. For example, human ferritin monomers have a mass of 20–21 kDa and a size of ~4 nm, whereas the 24-mer has a mass of ~400 kDa and a diameter of ~10 nm [34]. In addition, the free monomers have a flattened shape whereas the oligomer is a hollow sphere. Similarly, monomers of antibodies are composed of globular domains (e.g., immunoglobulin gamma heavy chain constant region 2B; UniProt ID P01867) but assemble into a Y-shaped tetramer, and in some cases higher-order structures with additional components. Even without considering oligomerization, the tertiary structure of proteins can adopt diverse shapes, including dense spheres (e.g., bovine chymotrypsinogen A; UniProt ID P00766), barrels (e.g., red fluorescent protein drFP583; UniProt ID Q9U6Y8), rods (e.g., human fibrinogen; UniProt ID P02675) and hooks (e.g., integrin α-E; UniProt ID P38570). Surface properties are also affected by oligomerization. For example, the (structure-based) isoelectric points of the tobacco RuBisCO large and small subunits of are 5.60 and 6.51 respectively, whereas the isoelectric point of the hetero-16-mer is 6.21.

Post-translational modifications can also alter both the shape and surface properties of proteins. Glycosylation significantly increases the mass of a protein (e.g., by ~1.9 kDa per site or 5–10% of a protein [35]), and affects the surface properties in a number of ways. On one hand, carbohydrates can shield the surface and prevent certain areas interacting with other molecules, for example to modify or suppress recognition by the immune system [36]. On the other hand, many carbohydrates feature charged monomers that alter the protein surface charge (distribution) [37]. Other post-translational modifications may be smaller but can nevertheless alter the surface properties of a protein dramatically. For example, a phosphate group has a diameter of less than 0.4 nm [38], but adds two negative charges (depending on the pH of the surrounding medium). Similarly, methylation and acetylation remove a potential hydrogen bonding site and increase the hydrophobicity of the surface [39]. Protein databases such as UniProt (https://www.uniprot.org/) provide annotated information about potential and reported post-translational modifications, but the set of modifications present on a given protein may vary and is difficult to predict a priori, especially if a protein is produced in a heterologous host that may not create native modifications authentically. For example, human proteins expressed in yeast often contain high-mannose glycans lacking the charged, terminal *N*-acetylneuraminic acid residues normally found in the native protein [40]. This makes it difficult to account for post-translational modifications when calculating descriptors, especially if samples of purified, authentic protein are not available for analysis. Nevertheless, all these layers of structural complexity should be considered when designing and calculating descriptors of protein surface and protein shape properties in order to obtain values that are representative of the actual protein.

3. Descriptor availability, redundancy and applicability

Given the structural complexity of proteins, many descriptors have been developed to cover different properties. Some descriptors were designed for the characterization of small molecules and capture atomic properties such the number of specific nitrogen atoms or chemical groups such as aromatic rings [21]. These are often of limited use for proteins because they are unlikely to capture the surface and shape properties in a meaningful way and their values can be similar even for proteins lacking structural homology (see the carbon atom example above). Others have been developed specifically to quantify protein shape and surface properties, such as the number, shape and size of surface patches (contiguous protein surface areas with a consistent property, like hydrophobicity or positive charge) [41,42]. However, the underlying property (e.g., positive charge) is typically a continuous parameter and it is difficult to define the boundaries and thus the size and number of patches. A given position on the protein surface is unlikely to have an exactly zero net charge but will instead exhibit a slight charge due to the combined effects of positively and negatively charged amino acid side chains near that position [43]. A common workaround is to define thresholds for the patch boundaries, for example the minimal positive charge could be 10 eV. However, this means the boundary condition of a patch is defined arbitrarily, and the relevance of the threshold with respect to applications such as protein binding to a charged chromatography ligand would be unclear [24,44]. This is often addressed by creating sets of similar descriptors that use different thresholds, e.g. the k largest patches. But even this creates new difficulties because introducing sets of similar descriptors substantially inflates the total number of descriptors. Furthermore, the descriptors within such a set are highly redundant and collinear by design, aggravating the selection of the most relevant descriptors during subsequent QSAR model building [45]. An informed threshold selection and/or descriptor pre-selection by experts with domain knowledge (e.g., in chromatography) can reduce the number of descriptors but will rely on personal experience and may be unintentionally biased. Therefore, a more rational definition of thresholds and descriptors would be helpful in this context as was also recently discussed in the context of small molecules and when comparing self-supervised and manually selected features [46,47].

Another major limitation of many protein descriptors is their averaging effect. For example, calculating the dipole momentum of a protein will integrate polarity over the entire molecule, which can mask distinct differences both on the surface and between surfaces. Indeed, molecules with different partial charge distributions may have the same dipole momentum (Fig. 1). Accordingly, the descriptor will not properly capture the intramolecular heterogeneity. The inability of Pearson's correlation coefficient to distinguish between different types of (non-randomly) scattering data can be regarded as an analogy in this context. Therefore, novel descriptors may be required that maintain information about the heterogeneity of properties such as shape and charge distribution.

The various types of descriptors can be calculated based on either the amino acid sequence or the protein structures such as those from the Protein Data Bank (PDB) using many different commercial or free-to-use software tools (Table 1). Some of these tools have recently been reviewed [45], and tool selection may be driven by the subsequent application of the descriptors and resulting QSAR model (e.g., the prediction of protein separation or solubility [48,49]).

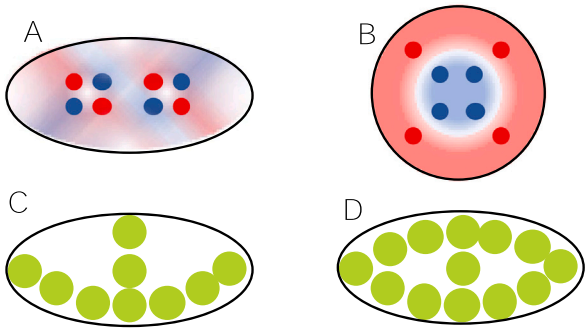


Fig. 1. Descriptor ambiguity. A. Schematic representation of a protein (black ellipse) with positive (red) and negative (blue) surface charges creating a net zero dipole moment. B. Alternative schematic for a different protein (black circle) with the same net zero dipole moment as in A but different surface charge distribution, i.e. a central patch of negative surface charge. C. Schematic representation of protein domains (green dots) and the corresponding minimum volume-enclosing ellipsoid use to calculate eccentricity (black ellipse). D. Alternative protein with the same minimum volume-enclosing ellipsoid as in C, but substantially different domain architecture and mass.

Regardless of which software tool is used to calculate the descriptors, the values can vary substantially based on the hydrogen atom addition and energy minimization operations applied to the structures before descriptor calculation. For example, we obtained electric dipole moments of catalase (UniProt ID P00432, pdb ID 3re8) ranging from 371 to 812 Debye (1.237×10^{-27} to 2.708×10^{-27} C m) when using Molecular Operating Environment (MOE; Chemical Computing Group, Canada) for the calculation. In contrast, an open access server provided a value of 121 Debye (0.404×10^{-27} C m) [70]. At best, this variability can be accommodated by calculating the descriptors of a protein several times for structures derived from individual hydrogen atom addition and energy minimization procedures or using slightly different parameters for the latter (e.g., force field, pH and temperature).

4. Protein shape descriptors

Protein shape descriptors often simplify the overall complexity of a protein structure by reporting sum parameters such as the molecular mass, the van der Waals (vdW) area or (S)ASA and corresponding volumes [71]. Similarly, protein properties may be averaged and represented through convex hulls [72] or equivalent spheres that share a certain feature with the protein of interest. Accordingly, the radii of such spheres are descriptors too: for example, the hydrodynamic radius (also called the Stokes radius, equivalent to diffusion properties, [73]) or the radius of gyration (equivalent to the moment of inertia, [74,75]), which can be used to predict polymer-induced precipitation and/or diffusion in porous media as well as potential sites of molecular interactions [76–79]. Recently, models have been developed that correlate the hydrodynamic radius and radius of gyration for unstructured proteins [80]. Furthermore, the radii of gyration and eccentricity are measures of compactness [19,81], and thus indicate how well a protein complies with the frequent assumption of a sphere-like shape made by many predictive models. Indeed, predictions about friction or sedimentation often assume smooth spheres [82] and so do models predicting unstructured regions of proteins. Eccentricity and the principal axes of inertia are interesting indicators of molecular anisotropy and are linked to ligand-binding sites [76]. They can mask distinct structural differences between proteins because molecules differing in shape may have the same eccentricity. Specifically, eccentricity is based on the minimum volume-enclosing ellipsoid [19], which can be identical even between proteins that occupy substantially different fractions of the volume of that ellipsoid (Fig. 1).

Shape descriptors can also account for the number of secondary structures and their relative abundance in a protein. This is important because motifs featuring, for example, up to four α-helices often mediate protein–protein interactions [83,84]. The descriptor calculation typically uses propensities derived from theoretical considerations or generalized experimental observations, as proposed by Chou and Fasman [85], and can be improved if data from actual protein three-dimensional structures are available from X-ray crystallography, nuclear magnetic resonance spectroscopy or homology modeling studies. Furthermore, the relative orientation of secondary structures, which is important for the formation of binding motifs, is not captured by these descriptors.

The protein shape is also affected by intrinsically unstructured, disordered or natively unfolded regions, which facilitate flexible and context-specific binding [86,87]. These are described as intrinsically disordered regions (IDRs) if they are found within an otherwise structured protein, or intrinsically disordered proteins (IDPs) if the entire protein is affected. IDRs can be predicted based on propensities similar to those of helices and sheets, for example using neural networks [87,88]. Accordingly, the corresponding descriptors share similar limitations such as the reliance on biased datasets over-representing globular proteins while neglecting other shapes [89].

At the level of primary structure, proteins can be described using graph theory approaches, such as the connectivity index and Wiener index. These indices facilitate primary sequence alignments [90] as well as the design of protein interaction networks [91,92], but are more likely to be relevant for small molecules. For example, the Wiener index is the sum of the shortest path between all pairs of vertices (e.g., atoms on the protein surface), yet the number of such pairs m increases as a function of the number of vertices n according to the equation $m = n \times (n - 1)/2$, so that values can be large for proteins but differences between proteins may be small despite substantial differences in shape. Interestingly, protein size and oligomerization, as well as structural elements such as linker sequences, seem to correlate with the relative abundance of amino acids like alanine, which may also be useful as a descriptor [93,94].

In contrast, global, rotation-invariant descriptors of protein shape may be derived from a series expansion of three-dimensional functions. Although these descriptors are based on the global shape of a protein, they account for distinct local features and therefore circumvent some of the shortcomings described above for conventional descriptors. One successful example is the so-called Zernike 3D descriptor, which allows the rapid identification of similar protein shapes and surfaces involved in protein–protein interactions [19,95,96]. Zernike 3D descriptors are vectors of, for example, 121 entries. Each entry represents a norm of a vector of coefficients derived from the expansion series of a three-dimensional function formulated on an orthogonal basis [96], and different three-dimensional functions can be used [95]. Accordingly, protein shape similarity can be quantified by calculating the difference in the Euclidian norms of the Zernike 3D descriptors of two proteins or using similar distance measures such as the Manhattan metric. In addition to this tensor algebra-based approach, linear algebra and other forms can be used as well [97]. The topic of 3D molecular descriptors for proteins has recently been applied and reviewed elsewhere [97–99]. In this context, the Laplace-Beltrami operator may be used as an alternative option to describe protein shapes invariant to rotation and translation [100,101]. Specifically, this operator can be used to calculate local geometric descriptors, such as the heat kernel signature [102], which can be used to create a fingerprint-like map for each node on a discretized, polygonal surface, for example of a protein. A challenge of this approach is the definition of meaningful time points at which a surface is to be evaluated, potentially creating a large number of (partially) redundant descriptors. There are various methods to reduce the dimensionality of the high-dimensional local descriptors and condense them into predefined global

Table 1
Overview of software packages used for descriptor calculation. A similar overview has been published before [45].

Application	Tool name	Type of descriptors	Input format	Output	Software Type	Open access	Web address	Ref.
General	CDK	Topological, constitutional, geometric, electronic, hybrid descriptors, fingerprints (ECFP, Daylight, MACCS, and others)	SMILES, SDF, InChI, MOL2, CML, and others	One or more values of varying types (Boolean, Double, Integer)	Java Library	yes	http://cdk.github.io	[50]
	MOE	2d, 3d descriptors, topological, physical properties, structural keys	PDB, SDF, SMILES, and others	n.a.	Software	no	www.chemcomp.com	n.a.
	PyDPI	Amino acid, di-, and tri-peptide composition, autocorrelation descriptors, CTD descriptors, sequence order coupling, quasi-sequence order descriptors, pseudo amino acid composition, descriptors, amphiphilic pseudo amino acid composition descriptors, conjoint triad features	n.a.	Binary, integer and continuous scalars, group of scalars descriptors can be interpreted as 1D, 2D, or 3D arrays**	Python toolkit	yes	https://pypi.org/project/pydpi/#description	[51]
	Rcpi	Amino acid composition, autocorrelation, CTD descriptors, quasi-sequence order, pseudo-amino acid composition, PSSM profile, molecular descriptors (2D/3D), including constitutional, topological, geometrical, electronic descriptors, molecular fingerprints (PCM, GO similarity, sequence similarity)	SMILES, DF files, FASTA, PDB	n.a.	R package	yes	https://bioconductor.org/packages/release/bioc/vignettes/Rcpi/inst/doc/Rcpi.html; http://research.chem.psu.edu/pcjgroup/adapt.html	[52]
Small molecules	ADAPT 260 *	Topological, geometrical, electronic	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
	ADMET predictor	Constitutional, functional group counts, topological, E-state, 3d descriptors, molecular patterns, acid-base ionization, empirical estimates of quantum	SDF	n.a.	software	no	www.simulations-plus.com	n.a.
	ADRIANA. Code	Constitutional, functional group counts, Lipinski rule-of-five, topological, E-state, Moriguchi, Meylan flags, 3d descriptors, molecular patterns	n.a.	n.a.	software	no	http://www.akosgmbh.de/molnet/adriana.htm	n.a.
	ALOGPS2.1 *	logP, logS	n.a.	n.a.	n.a.	yes	www.vcclab.org	n.a.
	BioChem	Constitution, topology, connectivity, kappa, estate, autocorrelation, molecular properties, charge, moe-type descriptors	SMILES, SDF, SMI-file	Integer and continuous scalars, groups can be interpreted as arrays**	Web-server	yes	http://biotriangle.scbdd.com	[53]
	BlueDesc	Constitutional descriptors, topological descriptors, connectivity indices, 2d-autocorrelation, topological charges indices, geometrical descriptors, WHIM descriptors, functional groups, molecular properties	3D-structures	Integer and continuous scalars, groups can be interpreted as arrays**	Java-Tool	yes	http://www.ra.cs.uni-tuebingen.de/software/bluedesc/welcome_e.html	n.a.
	ChemDes	Molecular descriptors	SMILES, SDF, SMI-file	Integer and continuous scalars, groups can be interpreted as arrays**	web-based platform	yes	www.scbdd.com/chemdes	[54]
	Chemopy	Constitution, topology, connectivity, E-state, kappa, Basak, Burden, autocorrelation (Moreau-Broto, Moran, Geary), charge, property, moe-type, geometric, CPSA, RDF, MoRSE, WHIM, fingerprints	SMILES, MOL	Integer and continuous scalars, groups can be interpreted as arrays**	Python Package	free	https://code.google.com/archive/p/pychem/downloads	[55]
	CODESSA	Constitutional, topological geometrical, charge-related, semi-empirical, thermodynamical	n.a.	Numerical characteristics	Software	no	www.codessa-pro.com	n.a.
	DRAGON	Constitutional, topological 2d-autocorrelations, geometrical, WHIM, GETAWAY, RDF, functional groups, etc.	MDL, Sybyl, SMILES, CML, and others	Integer and continuous scalars, groups can be interpreted as arrays**	Software	no	www.talete.mi.it	n.a.
	E-DRAGON*	Molecular descriptors	n.a.	n.a.	Software	yes	www.vcclab.org/lab/edragon/	n.a.

(*continued on next page*)

Table 1 (*continued*)

Application	Tool name	Type of descriptors	Input format	Output	Software Type	Open access	Web address	Ref.
	JOELib	Counting, topological, geometrical properties	SMILES, SDF, CML, CTX, FLAT, MOL2, and others (PDB in progess)	n.a.	Java-Package	yes	http://www.ra.cs.uni-tuebingen.de/software/joelib/introduction.html	n.a.
	KRAKENX	Quantum chemical, atomic charge-based, CPSA, EVA/EEVA, shape, geometry, topographical, Coulomb matrix, graph energy	SDF	Integer and continuous scalars, groups can be interpreted as arrays**, vectors, matrices	Software	yes	https://gitlab.com/vishsoft/krakenx	[56]
	MODEL*	Molecular descriptors	n.a.	n.a.	Web-server	yes	http://jing.cz3.nus.edu.sg/cgi-bin/model/model.cgi	n.a.
	MOLCONN-Z	Topological	SMILES, SDF, MOL2	Integer and continuous scalars, groups can be interpreted as arrays**	Software	no	www.edusoft-lc.com/molconn	n.a.
	MOLD2	1d, 2d	n.a.	Integer and continuous scalars, groups can be interpreted as arrays**	Software	yes	https://www.fda.gov/science-research/bioinformatics-tools/mold2	[57]
	MOLGEN-QSPR	Constitutional, topological, geometrical	SDF	n.a.	Software	no	www.molgen.molgenqspr.html	n.a.
	Mordred	2d, 3d	SMILES, SDF	Integer and continuous scalars, boolean, groups can be interpreted as arrays**	Python Package	yes	https://github.com/mordred-descriptor/mordred	[21]
	OEChem TK	166-bit MACCS, LINGO, circular, path (Daylight-like)	FASTA, SMILES, PDB, andothers	n.a.	programming library	no	www.eyesopen.com	n.a.
	OpenBabel	MOLPRINT2D, 166-bit MACCS, Daylight fingerprint (FP2), structural key fingerprints	SMILES, PDB, MOL2, CML., InChl, and other (84 formats)	Numerical and textual descriptors	Python Modul / Software	yes	www.openbabel.org	[58]
	PADEL*	1d, 2d, 3d descriptors, molecular fingerprints	n.a.	1d, 2d, and 3d descriptors	n.a.	yes	www.padel.nus.edu.sg	[59]
	PowerMV	Constitutional, atom pairs, fingerprints, BCUT	MDL, SDF	Bit strings, continuous, collections, and counts	Software	yes	www.niss.org/PowerMV	[60]
	PreADMET	Constitutional, topological, geometrical, physicochemical, etc.	n.a.	n.a.	Web-server	yes	http://preadmet.bmdrc.org	n.a.
	QikProp	Topological descriptors, Mopac Descriptors, QikProp	n.a.	n.a.	Software	no	https://www.schrodinger.com/products/qikprop	n.a.
	QuBiLS-MAS	Topological molecular descriptors based	MOL, SDF, MDL	Continuous scalars, groups can be interpreted as arrays**	Software	yes	http://tomocomd.com/software/qubils-mas	[22]
	RDKit	2d, 3d descriptors, fingerprints	SDF, MOL2, PDB, and others	Integer and continuous scalars, vectors, dict	Toolkit	yes	https://www.rdkit.org	n.a.
	Sarchitect*	Constitutional, 2d, 3d	n.a.	n.a.	n.a.	no	www.strandls.com/sarchitect/index.html	n.a.
	VolSurf+	ADME relevant descriptors	n.a.	Quantitative numerical descriptors	Software	no	https://www.moldiscovery.com/software/vsplus/	n.a.
Protein	AAindex	No direct descriptor	n.a.	Integer and continuous scalars, group of scalars descriptors can be interpreted as 1d, 2d, or 3d arrays**	Database of amino acid indices	yes	https://pypi.org/project/pydpi/#description	[61]
	BioProt	Amino acid, di-, and tri-peptide composition, CTD descriptors, M-B, Moran, Geary autocorrelation, conjoint triad features, quasi-sequence order descriptors, sequence order coupling number, pseudo amino acid composition 1 / 2	raw aa sequence, FASTA	Integer and continuous scalars, group of scalars descriptors can be interpreted as 1d, 2d, or 3d arrays**	Web-server	yes	http://biotriangle.scbdd.com	[53]
	DescribePROT	Database of 13 complementary descriptors for 83 complete proteomes	n.a.	Binary, integer and continuous scalars, group of scalars can be interpreted as arrays**	Database of descriptors	yes	http://biomine.cs.vcu.edu/servers/DESCRIBEPROT/main.php	[62]
	iFeature	Amino acid composition, grouped amino acid composition, CTD descriptors, M-B, Moran,	FASTA	n.a.	Python Package / Web-server	yes	https://ifeature.erc.monash.edu	[63]

(*continued on next page*)

Table 1 (*continued*)

Application	Tool name	Type of descriptors	Input format	Output	Software Type	Open access	Web address	Ref.
		Geary autocorrelation, conjoint triad, quasi-sequence-order, pseudo-amino acid composition, PseKRAAC, AAindex, BLOSUM62, Z-scale						
	MuLiMs-MCoMPAs*	n.a.	n.a.	n.a.	n.a.	yes	http://mumbailocalinference.net/MCoMPAS/MCoMPAS.html	[64]
	ProFeat-SS*	n.a.	n.a.	n.a.	n.a.	yes	http://bidd.cz3.nus.edu.sg/software/ProFeat-SS/*	n.a.
	PROFEAT*	n.a.	n.a.	n.a.	n.a.	yes	http://bidd.cz3.nus.edu.sg/software/PROFEAT/*	n.a.
	ProPred*	n.a.	n.a.	n.a.	n.a.	yes	https://web.kuicr.kyoto-u.ac.jp/supp/propred/	n.a.
	ProtDCal*	n.a.	PDB, FASTA	n.a.	n.a.	yes	https://webs.iiitd.edu.in/raghava/protdcal/	[65]
	PROTEIN RECON*	n.a.	n.a.	n.a.	n.a.	yes	http://bioinf.modares.ac.ir/software/protein-recon/	n.a.
	Protparam	Molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index, grand average of hydropathicity (GRAVY)	raw aa sequence, Swiss-Prot/TrEMBL accession number (AC)	Integer and continuous scalars, group of scalars descriptors can be interpreted as 1d, 2d, or 3d arrays**	Web-server	yes	https://web.expasy.org/protparam/	[66]
	protr	Amino acid composition, grouped amino acid composition, CTD descriptors, M-B, Moran, Geary autocorrelation, conjoint triad, sequence order coupling number, quasi-sequence-order, pseudo-amino acid composition, amphiphilic pseudo-amino acid composition, pseudo amino acid compositions	raw aa sequence, FASTA	Integer and continuous scalars, group of scalars descriptors can be interpreted as 1d, 2d, or 3d arrays**	R Package	yes	https://nanx.app/protr/	[67]
	PseAAC	n.a.	FASTA	Integer and continuous scalars, group of scalars descriptors can be interpreted as 1d, 2d, or 3d arrays**	Web-server	yes	http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/	[68]
	PyBioMed (PyProtein)	Amino acid, di-, and tri-peptide composition, CTD descriptors, M-B, Moran, Geary autocorrelation, conjoint triad features, quasi-sequence order descriptors, sequence order coupling number, pseudo amino acid composition 1 / 2	PDB ID, UniprotID	Integer and continuous scalars, group of scalars descriptors can be interpreted as 1d, 2d, or 3d arrays**	Python Package	yes	https://pybiomed.readthedocs.io/en/latest/	[69]

*the websites were not accessible during the time of writing, which is why little information was retrievable; ** some descriptors represent amino acid or heavy atom counts, which can be regarded as sets of individual descriptors or as a single 1D-array. Additionally, there are descriptors that count pairs and triples, which can be represented as 2D or 3D-arrays accordingly.

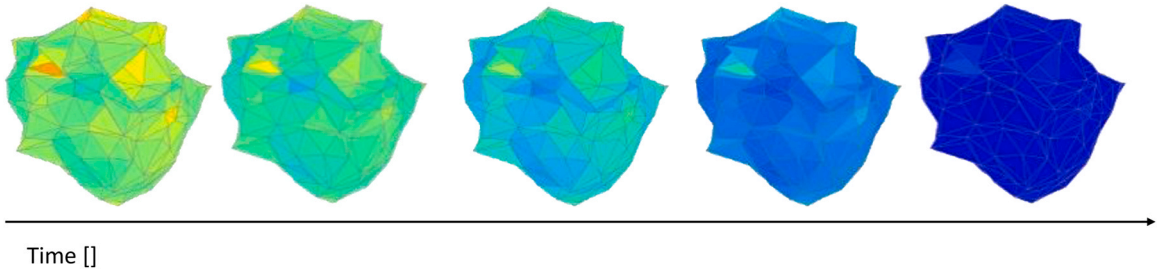


Fig. 2. Discretized polygonal surface of plastocyanin B'/B'' (UniProt ID P35477) colored according to its heat kernel signature [102] at various time points. The heat kernel signature of each vertex reflects a theoretical heat dissipation over time and serves as a tool to elucidate distinct local geometry determined by the intrinsic properties of the protein surface that can be used as a descriptor.

descriptors, e.g. by calculating the covariance matrix or aggregating eigenfunctions [101,103]. When combining Laplace-Beltrami operator-derived descriptors with methods such as bag-of-features [104], it may become possible to simultaneously capture protein shape and surface properties, for example by correlating protein properties with a heat kernel signature [102]. (Fig. 2).

5. Protein surface property descriptors

The functionality of the protein surface comprises various important properties that affect solubility and interactions with other molecules, including proteins, substrates, or ligands during chromatographic separation. These surface properties are defined by the functional groups of amino acid side chains exposed on the protein surface, including charged, polar or hydrophobic residues (among other classifications [105]), and may constitute patches on the surface if several amino acids of the same type cluster together [12]. The corresponding protein surface regions may partially overlap, making it difficult to define boundaries as discussed above. Protein surface descriptors typically report properties as counts (e.g., number of positively charged surface patches) or sums (e.g., total positively charged surface area), both of which can be normalized against another property such as the (S)ASA. Sets of descriptors can capture the same property using different thresholds (also known as gates), defined by a minimum number, size, or other parameter. Like shape descriptors, surface descriptors can capture properties globally and locally as discussed below for charged, polar and hydrophobic descriptors.

5.1. Charge

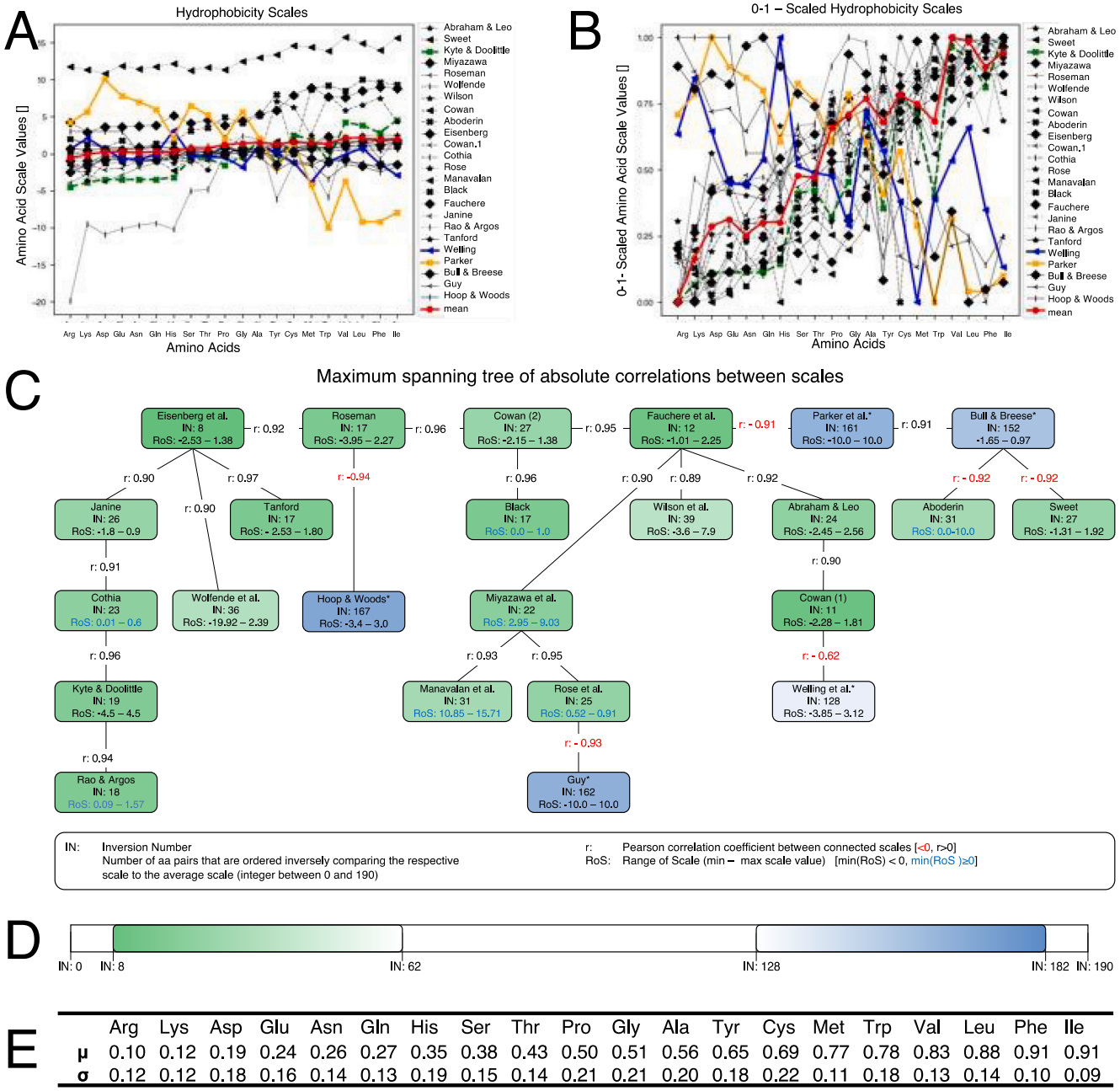
A prominent global descriptor of a protein's surface charge properties is the isoelectric point (pI). This can be calculated from the primary sequence alone [106] or by accounting for structural information reflecting the functional groups accessible on the protein surface [107]. Databases with proteome-wide isoelectric point information are now available [108] and this metric can be used, for example, as a predictor of crystallization efficiency [109] whereas it may be insufficient to predict protein–ligand interactions in a chromatography setting [110]. A property related to the isoelectric point is the overall net charge of a protein in a buffer with a given pH. In molecular dynamics simulations, this property correlates well with the colloidal interaction strength between proteins [111].

Protein surface charge can also be characterized by the number and size of positively and/or negatively charged patches, and thresholds for the minimum patch size (e.g., in $Å^2$ or nm^2) and charge (e.g., in eV) can be applied. Additional constraints may be formulated to specify the surface charge descriptors. For example, the charged surface area may be limited to the *k* largest patches or counting can be restricted to certain protein domains or parts thereof, such as the complementarity determining regions (CDRs) of an antibody. Such charge-based descriptors have been used in QSAR models for the successful prediction of retention on different types of charged chromatography resin [24,48]. Locally, descriptors may report residue pKa values, which can be combined with a protein's pseudo amino acid composition to discriminate between mesophilic and thermophilic proteins [112].

Other descriptors relate surface charge to the behavior of proteins in suspension. Specifically, the ζ-potential is the electric potential of a protein at its slipping plane when moving through a suspension. The ζ-potential can be calculated for different conditions, such as bulk suspensions or confined environments like pores [113,114], and depends on properties such as pH and conductivity, which must be specified accordingly [115]. The ζ-potential can be used to predict the likelihood of protein aggregation, with values close to zero indicating a high propensity and values < –40 or > 40 mV typically indicating colloidal stability [116]. A closely related descriptor is the Debye (screening) length (κ^{-1}), which is a measure of how quickly (in terms of distance) the effects of the electric potential of a protein decline in the surrounding suspension. Accordingly, the Debye length depends on the salt type and concentration as well as temperature, and can be used to describe the effect of suspension properties on the formation of a protein–polyelectrolyte complex [117].

5.2. Polarity

Whereas charge is defined by the presence of ions of opposing charge, polarity concerns partial charges that arise when the electronegativity differs between two covalently bound atoms [118]. Accordingly, polarity-based descriptors may account for surface properties based on charge, non-charged but polar features, or a combination of both, blurring the boundary between charge and polarity-based descriptors. Dipole moments exemplify a family of global, polarity-based descriptors, capturing the overall anisotropy of the surface (partial) charge distribution of a protein. On one hand, this family consists of the dipole direction and dipole moment of a protein, the latter distinguishing between proteins that are likely to precipitate in the presence of either ammonium sulfate or polyethylene glycol [18]. The (charge) separation distance and (charge) shape regularity are descriptors with a similar information content, which also applies to the ζ-dipole moment derived from the ζ-potential described above. On the other hand, dipole moments may also be calculated for each of the three axes of a Cartesian coordinate system or for the principal axes of inertia of the protein as discussed above. More selective measures of polarity can also be applied, for example by calculating the charge symmetry of variable regions of antibodies, providing useful descriptors for the *in silico* prediction of antibody solubility [25].



E

	Arg	Lys	Asp	Glu	Asn	Gln	His	Ser	Thr	Pro	Gly	Ala	Tyr	Cys	Met	Trp	Val	Leu	Phe	Ile
μ	0.10	0.12	0.19	0.24	0.26	0.27	0.35	0.38	0.43	0.50	0.51	0.56	0.65	0.69	0.77	0.78	0.83	0.88	0.91	0.91
σ	0.12	0.12	0.18	0.16	0.14	0.13	0.19	0.15	0.14	0.21	0.21	0.20	0.18	0.22	0.11	0.18	0.13	0.14	0.10	0.09

Fig. 3. Comparison of hydrophobicity scales. A. The hydrophobic propensities of the 24 hydrophobicity scales used in ProtScale are plotted for each of the 20 canonical amino acids. Mean (red), frequently referenced scales (Kyte & Doolittle, green), the scale ordered in reverse (Parker, orange), and the scale with the lowest absolute correlation to the other scales (Welling, blue) are highlighted in color. B. Same hydrophobicity scales as in A but with standardized values. C. Maximum spanning tree of the absolute correlations between hydrophobicity scales with inversion number (IN) to average scale and value range of scale (RoS). The r values at edges indicate the Pearson correlation coefficient between the connected scales. D. Color code of the boxes in C. according to the inversion number of the corresponding hydrophobicity scale. The inversion number quantifies the similarity of the order of amino acid hydrophobicity between the different hydrophobicity scale. Specifically, each scale is compared to the average amino acid order of hydrophobicity scaled between 0 and 1, e.g. an IN of 0 would indicate a perfect match with average whereas a value of n·(n-1)/2 (here:190) indicates a fully inverted order. E. Amino acid hydrophobicity averaged over all scales. The five inverted hydrophobicity scales (blue) were reverted when calculating the average to ensure a uniform direction in hydrophobicity. Amino acids are given in three-letter code in A, B and E.

The polarity of a protein may also be predicted from the primary sequence based on polarity propensities of the individual amino acids [119,120], but this approach neglects any structural information and is limited in value. The same applies to descriptors such as hydrophilicity indices, as well as the number of hydrogen-bond donors and acceptors, which are probably more relevant for small molecules because, for example, there are fewer donors per molecule compared to proteins, and differences thus facilitate better discrimination.

5.3. Hydrophobicity

Surface hydrophobicity is important for protein–protein recognition [121], aggregate formation [122], solubility, and proteolytic stability [123]. Many descriptors therefore encode hydrophobic properties. The total hydrophobicity of surface-exposed and solvent-accessible amino acid side chains is a global descriptor [124], which can be refined by applying a quadratic model to these values [125]. This modified descriptor was successfully used to predict protein

retention times during hydrophobic interaction chromatography. Similarly, partitioning coefficients, for example in octanol–water two-phase systems (logP(o/w)), can be used to describe the overall hydrophobicity/lipophilicity of a protein. However, such descriptors may be difficult to predict ab initio and the overall hydrophobicity does not reveal the distribution of hydrophobic patches on the protein surface. Similar to global charge descriptors, the hydrophobic dipole moment is a metric that quantifies the degree of asymmetry in surface hydrophobicity [126].

The hydrophobic dipole moment can also be calculated for fractions of a protein, such as a domain or secondary structure, as an indicator for intramolecular and intermolecular interactions [126]. However, information about the spatial distribution and relative orientation of the individual moments is not stored in these descriptors. Similarly, hydrophobicity can be calculated for the *k* largest patches, a certain number of patches, or a number of patches above a certain threshold, as discussed for surface charge above. It is important to note that if threshold selection is arbitrary and many thresholds are used, the number of collinear descriptors increases, which also increases the risk of overfitting QSAR models if the size of the training set is small compared to the number of descriptors [127].

At the primary sequence level, the hydrophobicity index of each amino acid in a protein can be aggregated to an overall hydrophobicity. However, many different definitions of this index have been developed over time [128], 24 of which are implemented in ProtScale, a protein identification and analysis tool on the Expasy server [129]. The definitions are based on either theoretical calculations or empirical data [130], such as the accessible surface of amino acids, the fraction of the number of different amino acids buried within proteins, the amino acid free energy of transfer from ethanol to water, and peptide retention times during reversed-phase chromatography [125]. Accordingly, the ranges and dimensions of hydrophobicity indices differ substantially. Nevertheless, 19 of the 24 indices in ProtScale are strongly connected, with an average Pearson's correlation coefficient of $r = 0.81$ (Fig. 3, Table S1). The remaining five indices still have an average Pearson's correlation coefficient of $r = 0.62$. They are negatively correlated with the first group of indices ($r = -0.72$), indicating an inversion of the direction of the scales compared to the other indices.

Ultimately, the selection of hydrophobicity descriptors will depend on the application. Whereas a dedicated tag-based hydrophobicity descriptor predicts the binding of tagged proteins during hydrophobic interaction chromatography more accurately than an index variant describing the entire surface, the accessible surface hydrophobicity can be used successfully to describe protein partitioning in aqueous two-phase systems [130].

5.4. Aggregation

Charge, polarity and hydrophobicity affect protein properties at the same time, and a holistic investigation should therefore account for combinations of these factors and their interactions. Calculating such combinations may be difficult a priori due to complex mutual interferences, including shape effects and environmental effects due to solvent and buffer properties. Therefore, a set of descriptors has been developed based on empirical polypeptide aggregation data to predict the tendency of proteins to form aggregates [131,132]. Although this approach can resolve differences caused by single amino acid substitutions, it depends on an empirical dataset for calibration and is therefore laborious to augment (including a lack of standardization between laboratories when generating such data) and is inherently biased in favor of polypeptides/proteins available in sufficient quantities. Alternative approaches include the Zyggregator family of descriptors based on more fundamental properties that can be updated quickly from the ever-growing protein structure databases, such as the α-helical propensities of amino acids [133]. Additional approaches to predict aggregation are statistical thermodynamic algorithms like TANGO [11], the spatial distribution of patches [12,111,134] or explicit molecular interactions [135].

6. Conclusion – need for and potentials of novel protein descriptors

Developing meaningful protein descriptors is a major challenge because it requires the integration of domain/application knowledge and has to balance information loss against redundancy and high dimensionality. Therefore, alternatives to commonly used aggregation functions such as mean, maxima, or thresholds have been proposed and include fuzzy measures to create global descriptors considering interdependencies for molecular descriptors [136,137]. These approaches seem promising in improving protein descriptors, especially as advanced structure prediction tools such as alpha-fold and ESM models provide access to an ever increasing number of protein structures [138,139].

Many descriptors currently used to capture protein shape and surface properties fail to account for the complexity of proteins, such as the relative position of charged surface patches or the anisotropy of properties in general. This can be a major limitation if distinct shape and surface properties strongly influence protein behavior in a given application. For example, close proximity between charged and hydrophobic surface patches can be required for protein binding to multimodal chromatography ligands [140], whereas the absolute number, size or strength of these patches may be much less important. Accordingly, descriptors should be cross-correlated and co-contextualized for each application using domain knowledge about the underlying molecular mechanisms. This approach helps to avoid excessive gating (i.e., the imposition of many threshold conditions for individual descriptors) and alternative descriptors can be defined on a rational basis. In this context, it can be valuable to use descriptors originating beyond the protein and biological molecule domain. For example, descriptors developed or first used for the analysis of cell images [141] or computer vision have been adapted to small molecules but not yet to proteins [103]. Three-dimensional functions such as the Laplace-Beltrami operator can also provide a good source of novel descriptors [142].

Ethics approval

Not applicable.

Consent for publication

All authors have seen a draft version of the manuscript and concur with its submission to the journal.

Code availability (software application or custom code)

Not applicable.

Funding

Not applicable.

CRediT authorship contribution statement

JB planned the study. JE and JB conducted the literature review, analyzed the data, prepared the figures and wrote the manuscript. JB revised the manuscript.

Data Availability

Data can be made available upon request to the corresponding author.

Acknowledgements

We wish to thank Dr. Richard M Twyman for editorial assistance. We thank Ronald Jäpel for supporting the calculation of protein dipole moments.

Conflicts of interest

The authors have no conflict of interest to declare.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.05.022.

References

[1] Ritala A, Häkkinen ST, Toivari M, Wiebe MG. Single cell protein-state-of-the-art, industrial landscape and patents 2001-2016. Front Microbiol 2017;8:2009. https://doi.org/10.3389/fmicb.2017.02009
[2] Hertzler SR, Lieblein-Boff JC, Weiler M, Allgeier C. Plant proteins: assessing their nutritional quality and effects on health and physical function. Nutrients 2020;12. https://doi.org/10.3390/nu12123704
[3] Charland N, Gobeil P, Pillet S, Boulay I, Séguin A, Makarkov A, Heizer G, Bhutada K, Mahmood A, Trépanier S, Hager K, Jiang-Wright J, Atkins J, Saxena P, Cheng MP, Vinh DC, Boutet P, Roman F, van der Most R, Ceregido MA, Dionne M, Tellier G, Gauthier J-S, Essink B, Libman M, Haffizulla J, Fréchette A, D'Aoust M-A, Landry N, Ward BJ. Safety and immunogenicity of an AS03-adjuvanted plant-based SARS-CoV-2 vaccine in adults with and without comorbidities. NPJ Vaccin 2022;7:142. https://doi.org/10.1038/s41541-022-00561-2
[4] Buyel JF, Bautista JA, Fischer R, Yusibov VM. Extraction, purification and characterization of the plant-produced HPV16 subunit vaccine candidate E7 GGG. J Chromatogr B 2012;880:19–26. https://doi.org/10.1016/j.jchromb.2011.11.010
[5] Shaaltiel Y, Gingis-Velitski S, Tzaban S, Fiks N, Tekoah Y, Aviezer D. Plant-based oral delivery of beta-glucocerebrosidase as an enzyme replacement therapy for Gaucher's disease. Plant Biotechnol J 2015;13:1033–40. https://doi.org/10.1111/pbi.12366
[6] Peters C, Brown S. Antibody-drug conjugates as novel anti-cancer chemotherapeutics. Biosci Rep 2015;35:1–20. https://doi.org/10.1042/BSR20150089
[7] Deng C, Huang T, Jiang Z, Lv X, Liu L, Chen J, Du G. Enzyme engineering and industrial bioprocess. Current Developments in Biotechnology and Bioengineering. Elsevier; 2019. p. 165–88.
[8] Exposito O, Bonfill M, Moyano E, Onrubia M, Mirjalili MH, Cusido RM, Palazon J. Biotechnological production of taxol and related taxoids: current state and prospects. Anti-Cancer Agent Me 2009;9:109–21. https://doi.org/10.2174/187152009787047761
[9] Roberts CJ. Protein aggregation and its impact on product quality. Curr Opin Biotechnol 2014;30:211–7. https://doi.org/10.1016/j.copbio.2014.08.001
[10] Conley GP, Viswanathan M, Hou Y, Rank DL, Lindberg AP, Cramer SM, Ladner RC, Nixon AE, Chen J. Evaluation of protein engineering and process optimization approaches to enhance antibody drug manufacturability. Biotechnol Bioeng 2011;108:2634–44. https://doi.org/10.1002/bit.23220
[11] van der Kant R, Karow-Zwick AR, van Durme J, Blech M, Gallardo R, Seeliger D, Assfalg K, Baatsen P, Compernolle G, Gils A, Studts JM, Schulz P, Garidel P, Schymkowitz J, Rousseau F. Prediction and reduction of the aggregation of monoclonal antibodies. J Mol Biol 2017;429:1244–61. https://doi.org/10.1016/j.jmb.2017.03.014
[12] Sankar K, Krystek SR, Jr, Carl SM, Day T, Maier JKX. AggScore: Prediction of aggregation-prone regions in proteins based on the distribution of surface patches. Proteins 2018;86:1147–56. https://doi.org/10.1002/prot.25594
[13] Möller J, Kuchemüller KB, Steinmetz T, Koopmann KS, Pörtner R. Model-assisted design of experiments as a concept for knowledge-based bioprocess development. Bioprocess Biosyst Eng 2019;42:867–82. https://doi.org/10.1007/s00449-019-02089-7
[14] Barley MH, Turner NJ, Goodacre R. Improved descriptors for the quantitative structure-activity relationship modeling of peptides and proteins. J Chem Inf Model 2018;58:234–43. https://doi.org/10.1021/acs.jcim.7b00488
[15] García-Jacas CR, Marrero-Ponce Y, Vivas-Reyes R, Suárez-Lezcano J, Martinez-Rios F, Terán JE, Aguilera-Mendoza L. Distributed and multicore QuBiLS-MIDAS software v2.0: Computing chiral, fuzzy, weighted and truncated geometrical molecular descriptors based on tensor algebra. J Comput Chem 2020;41:1209–27. https://doi.org/10.1002/jcc.26167
[16] Bertoni M, Duran-Frigola M, Badia-I-Mompel P, Pauls E, Orozco-Ruiz M, Guitart-Pla O, Alcalde V, Diaz VM, Berenguer-Llergo A, Brun-Heath I, Villegas N, de Herreros AG, Aloy P. Bioactivity descriptors for uncharacterized chemical compounds. Nat Commun 2021;12:3932. https://doi.org/10.1038/s41467-021-24150-4
[17] Todeschini R, Consonni V. Handbook of Molecular Descriptors. Wiley; 2000.
[18] Guo Y, Nishida N, Hoshino T. Quantifying the separation of positive and negative areas in electrostatic potential for predicting feasibility of ammonium sulfate for protein crystallization. J Chem Inf Model 2021;61:4571–81. https://doi.org/10.1021/acs.jcim.1c00505
[19] Han X, Sit A, Christoffer C, Chen S, Kihara D. A global map of the protein shape universe. PLoS Comput Biol 2019;15:e1006969. https://doi.org/10.1371/journal.pcbi.1006969
[20] Rostami R, Bashiri FS, Rostami B, Yu Z. A survey on data-driven 3D shape descriptors. Comput Graph Forum 2019;38:356–93. https://doi.org/10.1111/cgf.13536
[21] Moriwaki H, Tian Y-S, Kawashita N, Takagi T. Mordred: a molecular descriptor calculator. J Chemin- 2018;10:4. https://doi.org/10.1186/s13321-018-0258-y
[22] Valdés-Martiní JR, Marrero-Ponce Y, García-Jacas CR, Martinez-Mayorga K, Barigye SJ, Vaz d'Almeida YS, Pham-The H, Pérez-Giménez F, Morell CA. QuBiLS-MAS, open source multi-platform software for atom- and bond-based topological (2D) and chiral (2.5D) algebraic molecular descriptors computations. J Chemin- 2017;9:35. https://doi.org/10.1186/s13321-017-0211-5
[23] Lim H, Jeon H-N, Lim S, Jang Y, Kim T, Cho H, Pan J-G, No KT. Evaluation of protein descriptors in computer-aided rational protein engineering tasks and its application in property prediction in SARS-CoV-2 spike glycoprotein. Comput Struct Biotechnol J 2022;20:788–98. https://doi.org/10.1016/j.csbj.2022.01.027
[24] Robinson JR, Karkov HS, Woo JA, Krogh BO, Cramer SM. QSAR models for prediction of chromatographic behavior of homologous Fab variants. Biotechnol Bioeng 2017;114:1231–40. https://doi.org/10.1002/bit.26236
[25] Han X, Shih J, Lin Y, Chai Q, Cramer SM. Development of QSAR models for in silico screening of antibody solubility. MAbs 2022;14:2062807. https://doi.org/10.1080/19420862.2022.2062807
[26] Laurents DV, Subbiah S, Levitt M. Different protein sequences can give rise to highly similar folds through different stabilizing interactions. Protein Sci 1994;3:1938–44. https://doi.org/10.1002/pro.5560031105
[27] Pearson WR. An introduction to sequence similarity ("homology") searching. Curr Protoc Bioinforma Chapter 3 2013:3.1.1–8. https://doi.org/10.1002/0471250953.bi0301s42
[28] Koehl P, Levitt M. Sequence variations within protein families are linearly related to structural variations. J Mol Biol 2002;323:551–62. https://doi.org/10.1016/S0022-2836(02)00971-3
[29] Bernau CR, Knödler M, Emonts J, Jäpel RC, Buyel JF. The use of predictive models to develop chromatography-based purification processes. Front Bioeng Biotechnol 2022;10. https://doi.org/10.3389/fbioe.2022.1009102
[30] Comesana AE, Huntington TT, Scown CD, Niemeyer KE, Rapp VH. A systematic method for selecting molecular descriptors as features when training models for predicting physiochemical properties. Fuel 2022;321:123836. https://doi.org/10.1016/j.fuel.2022.123836
[31] Karlov DS, Sosnin S, Tetko IV, Fedorov MV. Chemical space exploration guided by deep neural networks. RSC Adv 2019;9:5151–7. https://doi.org/10.1039/c8ra10182e
[32] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, Peter Walter. Molecular Biology of the Cell. fourth ed. New York: Garland Science,; 2002.
[33] Ali MH, Imperiali B. Protein oligomerization: how and why. Bioorgan Med Chem 2005;13:5013–20. https://doi.org/10.1016/j.bmc.2005.05.037
[34] Knödler M, Opdensteinen P, Sankaranarayanan RA, Morgenroth A, Buhl EM, Mottaghy FM, Buyel JF. Simple plant-based production and purification of the assembled human ferritin heavy chain as a nanocarrier for tumor-targeted drug delivery and bioimaging in cancer therapy. Biotechnol Bioeng 2023;120:1038–54. https://doi.org/10.1002/bit.28312
[35] Gengenbach BB, Keil LL, Opdensteinen P, Muschen CR, Melmer G, Lentzen H, Buhrmann J, Buyel JF. Comparison of microbial and transient expression (tobacco plants and plant-cell packs) for the production and purification of the anticancer mistletoe lectin viscumin. Biotechnol Bioeng 2019;116:2236–49. https://doi.org/10.1002/bit.27076
[36] Boes A, Spiegel H, Edgue G, Kapelski S, Scheuermayer M, Fendel R, Remarque E, Altmann F, Maresch D, Reimann A, Pradel G, Schillberg S, Fischer R. Detailed functional characterization of glycosylated and nonglycosylated variants of malaria vaccine candidate PfAMA1 produced in Nicotiana benthamiana and analysis of growth inhibitory responses in rabbits. Plant Biotechnol J 2015;13:222–34. https://doi.org/10.1111/pbi.12255
[37] Hermentin P, Witzel R, Kanzy EJ, Diderrich G, Hoffmann D, Metzner H, Vorlop J, Haupt H. The hypothetical N-glycan charge: a number that characterizes protein glycosylation. Glycobiology 1996;6:217–30. https://doi.org/10.1093/glycob/6.2.217
[38] Kish MM, Viola RE. Oxyanion specificity of aspartate-beta-semialdehyde dehydrogenase. Inorg Chem 1999;38:818–20. https://doi.org/10.1021/ic981082j
[39] Kuczyńska-Wiśnik D, Moruno-Algara M, Stojowska-Swędrzyńska K, Laskowska E. The effect of protein acetylation on the formation and processing of inclusion bodies and endogenous protein aggregates in Escherichia coli cells. Microb Cell Factor 2016;15:189. https://doi.org/10.1186/s12934-016-0590-8
[40] Strasser R. Plant protein glycosylation. Glycobiology 2016;26:926–39. https://doi.org/10.1093/glycob/cww023

[41] Gamliel R, Kedem K, Kolodny R, Keasar C. A library of protein surface patches discriminates between native structures and decoys generated by structure prediction servers. BMC Struct Biol 2011;11:20. https://doi.org/10.1186/1472-6807-11-20
[42] Budowski-Tal I, Kolodny R, Mandel-Gutfreund Y. A novel geometry-based approach to infer protein interface similarity. Sci Rep 2018;8:8192. https://doi.org/10.1038/s41598-018-26497-z
[43] Hebditch M, Warwicker J. Web-based display of protein surface and pH-dependent properties for assessing the developability of biotherapeutics. Sci Rep 2019;9:1969. https://doi.org/10.1038/s41598-018-36950-8
[44] Dismer F, Hubbuch J. 3D structure-based protein retention prediction for ion-exchange chromatography. J Chromatogr A 2010;1217:1343–53. https://doi.org/10.1016/j.chroma.2009.12.061
[45] Danishuddin AUKhan. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. Drug Discov Today 2016;21:1291–302. https://doi.org/10.1016/j.drudis.2016.06.013
[46] García-González LA, Marrero-Ponce Y, Brizuela CA, García-Jacas CR. Overproduce and select, or determine optimal molecular descriptor subset via configuration space optimization? Application to the prediction of ecotoxicological endpoints. Mol Inform 2023:e2200227. https://doi.org/10.1002/minf.202200227
[47] García-Jacas CR, García-González LA, Martinez-Rios F, Tapia-Contreras IP, Brizuela CA. Handcrafted versus non-handcrafted (self-supervised) features for the classification of antimicrobial peptides: complementary or redundant. Brief Bioinform 2022;23. https://doi.org/10.1093/bib/bbac428
[48] Buyel JF, Woo JA, Cramer SM, Fischer R. The use of quantitative structure-activity relationship models to develop optimized processes for the removal of tobacco host cell proteins during biopharmaceutical production. J Chromatogr A 2013;1322:18–28. https://doi.org/10.1016/j.chroma.2013.10.076
[49] Rawi R, Mall R, Kunji K, Shen CH, Kwong PD, Chuang GY. PaRSnIP: sequence-based protein solubility prediction using gradient boosting machine. Bioinformatics 2018;34:1092–8. https://doi.org/10.1093/bioinformatics/btx662
[50] Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E. The chemistry development kit (CDK): an open-source java library for chemo- and bioinformatics. J Chem Inf Comp Sci 2003;43:493–500. https://doi.org/10.1021/ci025584y
[51] Cao D-S, Liang Y-Z, Yan J, Tan G-S, Xu Q-S, Liu S. PyDPI: freely available python package for chemoinformatics, bioinformatics, and chemogenomics studies. J Chem Inf Model 2013;53:3086–96. https://doi.org/10.1021/ci400127q
[52] Cao D-S, Xiao N, Xu Q-S, Chen AF. Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions. Bioinformatics 2015;31:279–81. https://doi.org/10.1093/bioinformatics/btu624
[53] Dong J, Yao Z-J, Wen M, Zhu M-F, Wang N-N, Miao H-Y, Lu A-P, Zeng W-B, Cao D-S. BioTriangle: a web-accessible platform for generating various molecular representations for chemicals, proteins, DNAs/RNAs and their interactions. J Chemin- 2016;8:34. https://doi.org/10.1186/s13321-016-0146-2
[54] Dong J, Cao D-S, Miao H-Y, Liu S, Deng B-C, Yun Y-H, Wang N-N, Lu A-P, Zeng W-B, Chen AF. ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation. J Chemin- 2015;7:60. https://doi.org/10.1186/s13321-015-0109-z
[55] Cao D-S, Xu Q-S, Hu Q-N, Liang Y-Z. ChemoPy: freely available python package for computational biology and chemoinformatics. Bioinformatics 2013;29:1092–4. https://doi.org/10.1093/bioinformatics/btt105
[56] Venkatraman V, Alsberg BK. KRAKENX: software for the generation of alignment-independent 3D descriptors. J Mol Model 2016;22:93. https://doi.org/10.1007/s00894-016-2957-5
[57] Hong H, Xie Q, Ge W, Qian F, Fang H, Shi L, Su Z, Perkins R, Tong W. Mold(2), molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. J Chem Inf Model 2008;48:1337–44. https://doi.org/10.1021/ci800038f
[58] O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open babel: an open chemical toolbox. J Chemin- 2011;3:33. https://doi.org/10.1186/1758-2946-3-33
[59] Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 2011;32:1466–74. https://doi.org/10.1002/jcc.21707
[60] Liu K, Feng J, Young SS. PowerMV: a software environment for molecular viewing, descriptor generation, data analysis and hit evaluation. J Chem Inf Model 2005;45:515–22. https://doi.org/10.1021/ci049847v
[61] Kawashima S, Kanehisa M. AAindex: amino acid index database. Nucleic Acids Res 2000;28:374. https://doi.org/10.1093/nar/28.1.374
[62] Zhao B, Katuwawala A, Oldfield CJ, Dunker AK, Faraggi E, Gsponer J, Kloczkowski A, Malhis N, Mirdita M, Obradovic Z, Söding J, Steinegger M, Zhou Y, Kurgan L. DescribePROT: database of amino acid-level protein structure and function predictions. Nucleic Acids Res 2021;49:D298–308. https://doi.org/10.1093/nar/gkaa931
[63] Chen Z, Zhao P, Li F, Leier A, Marquez-Lago TT, Wang Y, Webb GI, Smith AI, Daly RJ, Chou K-C, Song J. iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics 2018;34:2499–502. https://doi.org/10.1093/bioinformatics/bty140
[64] Contreras-Torres E, Marrero-Ponce Y, Terán JE, García-Jacas CR, Brizuela CA, Sánchez-Rodríguez JC. MuLiMs-MCoMPAs: a novel multiplatform framework to compute tensor algebra-based three-dimensional protein descriptors. J Chem Inf Model 2020;60:1042–59. https://doi.org/10.1021/acs.jcim.9b00629
[65] Romero-Molina S, Ruiz-Blanco YB, Green JR, Sanchez-Garcia E. ProtDCal-Suite: a web server for the numerical codification and functional analysis of proteins. Protein Sci 2019;28:1734–43. https://doi.org/10.1002/pro.3673
[66] Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A. Protein identification and analysis tools on the ExPASy Server. In: Walker JM, editor. The Proteomics Protocols Handbook. Totowa, NJ: Humana Press; 2005. p. 571–607.
[67] Xiao N, Cao D-S, Zhu M-F, Xu Q-S. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. Bioinformatics 2015;31:1857–9. https://doi.org/10.1093/bioinformatics/btv042
[68] Shen H-B, Chou K-C. PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. Anal Biochem 2008;373:386–8. https://doi.org/10.1016/j.ab.2007.10.012
[69] Dong J, Yao Z-J, Zhang L, Luo F, Lin Q, Lu A-P, Chen AF, Cao D-S. PyBioMed: a python library for various molecular representations of chemicals, proteins and DNAs and their interactions. J Chemin- 2018;10:16. https://doi.org/10.1186/s13321-018-0270-2
[70] Felder CE, Prilusky J, Silman I, Sussman JL. A server and database for dipole moments of proteins. Nucleic Acids Res 2007;35:W512–21. https://doi.org/10.1093/nar/gkm307
[71] Ali SA, Hassan MI, Islam A, Ahmad F. A review of methods available to estimate solvent-accessible surface areas of soluble proteins in the folded and unfolded states. Curr Protein Pept Sci 2014;15:456–76. https://doi.org/10.2174/1389203715666140327114232
[72] Fleming PJ, Fleming KG. HullRad: fast calculations of folded and disordered protein and nucleic acid hydrodynamic properties. Biophys J 2018;114:856–69. https://doi.org/10.1016/j.bpj.2018.01.002
[73] de Torre JLa García, Huertas ML, Carrasco B. Calculation of hydrodynamic properties of globular proteins from their atomic-level structure. Biophys J 2000;78:719–30. https://doi.org/10.1016/S0006-3495(00)76630-6
[74] Stepto R, Chang T, Kratochvíl P, Hess M, Horie K, Sato T, Vohlídal J. Definitions of terms relating to individual macromolecules, macromolecular assemblies, polymer solutions, and amorphous bulk polymers (IUPAC Recommendations 2014. Pure Appl Chem 2015;87:71–120. https://doi.org/10.1515/pac-2013-0201
[75] Funari R, Bhalla N, Gentile L. Measuring the radius of gyration and intrinsic flexibility of viral proteins in buffer solution using small-angle X-ray Scattering. ACS Meas Sci Au 2022;2:547–52. https://doi.org/10.1021/acsmeasuresciau.2c00048
[76] Foote J, Raman A. A relation between the principal axes of inertia and ligand binding. Proc Natl Acad Sci USA 2000;97:978–83. https://doi.org/10.1073/pnas.97.3.978
[77] He L, Niemeyer B. A novel correlation for protein diffusion coefficients based on molecular weight and radius of gyration. Biotechnol Prog 2003;19:544–8. https://doi.org/10.1021/bp0256059
[78] Sim S-L, He T, Tscheliessnig A, Mueller M, Tan RBH, Jungbauer A. Protein precipitation by polyethylene glycol: a generalized model based on hydrodynamic radius. J Biotechnol 2012;157:315–9. https://doi.org/10.1016/j.jbiotec.2011.09.028
[79] Maier RS, Schure MR. Transport properties and size exclusion effects in wide-pore superficially porous particles. Chem Eng Sci 2018;185:243–55. https://doi.org/10.1016/j.ces.2018.03.041
[80] Nygaard M, Kragelund BB, Papaleo E, Lindorff-Larsen K. An efficient method for estimating the hydrodynamic radius of disordered protein conformations. Biophys J 2017;113:550–7. https://doi.org/10.1016/j.bpj.2017.06.042
[81] Lobanov MY, Bogatyreva NS, Galzitskaya OV. Radius of gyration as an indicator of protein structure compactness. Mol Biol 2008;42:623–8. https://doi.org/10.1134/S0026893308040195
[82] Erickson HP. Size and shape of protein molecules at the nanometer level determined by sedimentation, gel filtration, and electron microscopy. Biol Proced Online 2009;11:32–51. https://doi.org/10.1007/s12575-009-9008-x
[83] Harrison RS, Shepherd NE, Hoang HN, Ruiz-Gómez G, Hill TA, Driver RW, Desai VS, Young PR, Abbenante G, Fairlie DP. Downsizing human, bacterial, and viral proteins to short water-stable alpha helices that maintain biological potency. Proc Natl Acad Sci USA 2010;107:11686–91. https://doi.org/10.1073/pnas.1002498107
[84] de Araujo AD, Lim J, Wu K-C, Hoang HN, Nguyen HT, Fairlie DP. Landscaping macrocyclic peptides: stapling hDM2-binding peptides for helicity, protein affinity, proteolytic stability and cell uptake. RSC Chem. Biol 3. 2022. p. 895–904. https://doi.org/10.1039/d1cb00231g
[85] Chou PY, Fasman GD. Prediction of protein conformation. Biochem-Us 1974;13:222–45. https://doi.org/10.1021/bi00699a002
[86] Covarrubias AA, Cuevas-Velazquez CL, Romero-Perez PS, Rendon-Luna DF, Chater CCC. Structural disorder in plant proteins: where plasticity meets sessility. Cell Mol Life Sci 2017;74:3119–47. https://doi.org/10.1007/s00018-017-2557-2
[87] Ferron F, Longhi S, Canard B, Karlin D. A practical overview of protein disorder prediction methods. Proteins 2006;65:1–14. https://doi.org/10.1002/prot.21075
[88] Orlando G, Raimondi D, Codicè F, Tabaro F, Vranken W. Prediction of disordered regions in proteins with recurrent neural networks and protein dynamics. J Mol Biol 2022;434:167579. https://doi.org/10.1016/j.jmb.2022.167579
[89] Mészáros B, Erdos G, Dosztányi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Res 2018;46:W329–37. https://doi.org/10.1093/nar/gky384

[90] Randić M. On the history of the connectivity index: from the connectivity index to the exact solution of the protein alignment problem. SAR QSAR Environ Res 2015;26:523–55. https://doi.org/10.1080/1062936X.2015.1076890
[91] Zhang P, Tao L, Zeng X, Qin C, Chen S, Zhu F, Li Z, Jiang Y, Chen W, Chen Y-Z. A protein network descriptor server and its use in studying protein, disease, metabolic and drug targeted networks. Brief Bioinform 2017;18:1057–70. https://doi.org/10.1093/bib/bbw071
[92] Munteanu CR, González-Díaz H, Borges F, de Magalhães AL. Natural/random protein classification models based on star network topological indices. J Theor Biol 2008;254:775–83. https://doi.org/10.1016/j.jtbi.2008.07.018
[93] Carugo O. Amino acid composition and protein dimension. Protein Sci 2008;17:2187–91. https://doi.org/10.1110/ps.037762.108
[94] Brüne D, Andrade-Navarro MA, Mier P. Proteome-wide comparison between the amino acid composition of domains and linkers. BMC Res Notes 2018;11:117. https://doi.org/10.1186/s13104-018-3221-0
[95] Daberdaku S, Ferrari C. Exploring the potential of 3D Zernike descriptors and SVM for protein-protein interface prediction. BMC Bioinforma 2018;19:35. https://doi.org/10.1186/s12859-018-2043-3
[96] Sael L, Li B, La D, Fang Y, Ramani K, Rustamov R, Kihara D. Fast protein tertiary structure retrieval based on global surface shape similarity. Proteins 2008;72:1259–73. https://doi.org/10.1002/prot.22030
[97] Marrero-Ponce Y, Teran JE, Contreras-Torres E, García-Jacas CR, Perez-Castillo Y, Cubillan N, Peréz-Giménez F, Valdés-Martini JR. LEGO-based generalized set of two linear algebraic 3D bio-macro-molecular descriptors: theory and validation by QSARs. J Theor Biol 2020;485:110039. https://doi.org/10.1016/j.jtbi.2019.110039
[98] Terán JE, Marrero-Ponce Y, Contreras-Torres E, García-Jacas CR, Vivas-Reyes R, Terán E, Torres FJ. Tensor algebra-based geometrical (3d) biomacro-molecular descriptors for protein research: theory, applications and comparison with other methods. Sci Rep 2019;9:11391. https://doi.org/10.1038/s41598-019-47858-2
[99] Marrero-Ponce Y, Contreras-Torres E, García-Jacas CR, Barigye SJ, Cubillán N, Alvarado YJ. Novel 3D bio-macromolecular bilinear descriptors for protein science: predicting protein structural classes. J Theor Biol 2015;374:125–37. https://doi.org/10.1016/j.jtbi.2015.03.026
[100] Caissard T, Coeurjolly D, Lachaud J-O, Roussillon T. Laplace–beltrami operator on digital surfaces. J Math Imaging Vis 2019;61:359–79. https://doi.org/10.1007/s10851-018-0839-4
[101] Wang Z, Lin H. 3D shape retrieval based on Laplace operator and joint Bayesian model. Vis Inform 2020;4:69–76. https://doi.org/10.1016/j.visinf.2020.08.002
[102] Kim S-G, Chung MK, Seo S, Schaefer SM, van Reekum CM, Davidson RJ. Heat kernel smoothing via laplace-beltrami eigenfunctions and its application to subcortical structure modeling. In: Ho Y-S, editor. Advances in Image and Video Technology. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. p. 36–47.
[103] Seddon MP, Cosgrove DA, Packer MJ, Gillet VJ. Alignment-free molecular shape comparison using spectral geometry: the framework. J Chem Inf Model 2019;59:98–116. https://doi.org/10.1021/acs.jcim.8b00676
[104] Litman R, Bronstein A, Bronstein M, Castellani U. Supervised learning of bag-of-features shape descriptors using sparse coding. Comput Graph Forum 2014;33:127–36. https://doi.org/10.1111/cgf.12438
[105] Li C, Li X, Lin Y-X. Numerical characterization of protein sequences based on the generalized Chou's pseudo amino acid composition. Appl Sci 2016;6:406. https://doi.org/10.3390/app6120406
[106] Sillero A, Ribeiro JM. Isoelectric points of proteins: theoretical determination. Anal Biochem 1989;179:319–25. https://doi.org/10.1016/0003-2697(89)90136-x
[107] Li H, Robertson AD, Jensen JH. Very fast empirical prediction and rationalization of protein pKa values. Proteins 2005;61:704–21. https://doi.org/10.1002/prot.20660
[108] Kozlowski LP. Proteome-pI 2.0: proteome isoelectric point database update. Nucleic Acids Res 2022;50:D1535–40. https://doi.org/10.1093/nar/gkab944
[109] Kantardjieff KA, Rupp B. Protein isoelectric point as a predictor for increased crystallization screening efficiency. Bioinformatics 2004;20:2162–8. https://doi.org/10.1093/bioinformatics/bth066
[110] Leisi R, Wolfisberg R, Nowak T, Caliaro O, Hemmerle A, Roth NJ, Ros C. Impact of the isoelectric point of model parvoviruses on viral retention in anion-exchange chromatography. Biotechnol Bioeng 2021;118:116–29. https://doi.org/10.1002/bit.27555
[111] Brunsteiner M, Flock M, Nidetzky B. Structure based descriptors for the estimation of colloidal interactions and protein aggregation propensities. PLoS ONE 2013;8:e59797. https://doi.org/10.1371/journal.pone.0059797
[112] Fan G-L, Liu Y-L, Wang H. Identification of thermophilic proteins by incorporating evolutionary and acid dissociation information into Chou's general pseudo amino acid composition. J Theor Biol 2016;407:138–42. https://doi.org/10.1016/j.jtbi.2016.07.010
[113] Vitarelli MJ, Talaga DS. Theoretical models for electrochemical impedance spectroscopy and local zeta-potential of unfolded proteins in nanopores. J Chem Phys 2013;139. https://doi.org/10.1063/1.4819470
[114] Hunter RJ. Zeta Potential in Colloid Science: Principles and Applications. Academic Press; 1981.
[115] Kamble S, Agrawal S, Cherumukkil S, Sharma V, Jasra RV, Munshi P. Revisiting zeta potential, the key feature of interfacial phenomena, with applications and recent advancements. ChemistrySelect 2022;7. https://doi.org/10.1002/slct.202103084
[116] Kumar A, Dixit CK. Methods for characterization of nanoparticles. Advances in Nanomedicine for the Delivery of Therapeutic Nucleic Acids. Elsevier; 2017. p. 43–58.
[117] Seyrek E, Dubin PL, Tribet C, Gamble EA. Ionic strength dependence of protein-polyelectrolyte interactions. Biomacromolecules 2003;4:273–82. https://doi.org/10.1021/bm025664a
[118] Jensen WB. The Origin of the "Delta" Symbol for Fractional Charges. J Chem Educ 2009;86:545. https://doi.org/10.1021/ed086p545
[119] Grantham R. Amino acid difference formula to help explain protein evolution. Science 1974;185:862–4. https://doi.org/10.1126/science.185.4154.862
[120] Zimmerman JM, Eliezer N, Simha R. The characterization of amino acid sequences in proteins by statistical methods. J Theor Biol 1968;21:170–201. https://doi.org/10.1016/0022-5193(68)90069-6
[121] Young L, Jernigan RL, Covell DG. A role for surface hydrophobicity in protein-protein recognition. Protein Sci 1994;3:717–29. https://doi.org/10.1002/pro.5560030501
[122] Alam P, Siddiqi K, Chturvedi SK, Khan RH. Protein aggregation: From background to inhibition strategies. Int J Biol Macromol 2017;103:208–19. https://doi.org/10.1016/j.ijbiomac.2017.05.048
[123] Murby M, Samuelsson E, Nguyen TN, Mignard L, Power U, Binz H, Uhlen M, Stahl S. Hydrophobicity engineering to increase solubility and stability of a recombinant protein from respiratory syncytial virus. Eur J Biochem 1995;230:38–44. https://doi.org/10.1111/j.1432-1033.1995.0038i.x
[124] Berggren K, Wolf A, Asenjo JA, Andrews BA, Tjerneld F. The surface exposed amino acid residues of monomeric proteins determine the partitioning in aqueous two-phase systems. Biochim Biophys Acta 2002;1596:253–68. https://doi.org/10.1016/S0167-4838(02)00222-4
[125] Lienqueo ME, Mahn A, Asenjo JA. Mathematical correlations for predicting protein retention times in hydrophobic interaction chromatography. J Chromatogr A 2002;978:71–9. https://doi.org/10.1016/S0021-9673(02)01358-4
[126] Eisenberg D, Weiss RM, Terwilliger TC, Wilcox W. Hydrophobic moments and protein structure. Faraday Symp Chem Soc 1982;17:109. https://doi.org/10.1039/FS9821700109
[127] Goodarzi M, Dejaegher B, Vander Heyden Y. Feature selection methods in QSAR studies. J AOAC Int 2012;95:636–51. https://doi.org/10.5740/jaoacint.SGE_Goodarzi
[128] Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105–32. https://doi.org/10.1016/0022-2836(82)90515-0
[129] Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 2003;31:3784–8. https://doi.org/10.1093/nar/gkg563
[130] Mahn A, Lienqueo ME, Salgado JC. Methods of calculating protein hydrophobicity and their application in developing correlations to predict hydrophobic interaction chromatography retention. J Chromatogr A 2009;1216:1838–44. https://doi.org/10.1016/j.chroma.2008.11.089
[131] Agrawal NJ, Kumar S, Wang X, Helk B, Singh SK, Trout BL. Aggregation in protein-based biotherapeutics: computational studies and tools to identify aggregation-prone regions. J Pharm Sci 2011;100:5081–95. https://doi.org/10.1002/jps.22705
[132] Conchillo-Solé O, de Groot NS, Avilés FX, Vendrell J, Daura X, Ventura S. AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. BMC Bioinforma 2007;8:65. https://doi.org/10.1186/1471-2105-8-65
[133] Tartaglia GG, Vendruscolo M. The Zyggregator method for predicting protein aggregation propensities. Chem Soc Rev 2008;37:1395–401. https://doi.org/10.1039/b706784b
[134] Woo J, Parimal S, Brown MR, Heden R, Cramer SM. The effect of geometrical presentation of multimodal cation-exchange ligands on selective recognition of hydrophobic regions on protein surfaces. J Chromatogr A 2015;1412:33–42. https://doi.org/10.1016/j.chroma.2015.07.072
[135] Heads JT, Lamb R, Kelm S, Adams R, Elliott P, Tyson K, Topia S, West S, Nan R, Turner A, Lawson ADG. Electrostatic interactions modulate the differential aggregation propensities of IgG1 and IgG4P antibodies and inform charged residue substitutions for improved developability. Protein Eng, Des Sel 2019;32:277–88. https://doi.org/10.1093/protein/gzz046
[136] García-Jacas CR, Cabrera-Leyva L, Marrero-Ponce Y, Suárez-Lezcano J, Cortés-Guzmán F, Pupo-Meriño M, Vivas-Reyes R. Choquet integral-based fuzzy molecular characterizations: when global definitions are computed from the dependency among atom/bond contributions (LOVIs/LOEIs). J Chemin-2018;10:51. https://doi.org/10.1186/s13321-018-0306-7
[137] García-Jacas CR, Cabrera-Leyva L, Marrero-Ponce Y, Suárez-Lezcano J, Cortés-Guzmán F, García-González LA. GOWAWA aggregation operator-based global molecular characterizations: weighting atom/bond contributions (LOVIs/LOEIs) according to their influence in the molecular encoding. Mol Inform 2018;37:e1800039. https://doi.org/10.1002/minf.201800039
[138] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, Dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, Rives A. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 2023;379:1123–30. https://doi.org/10.1126/science.ade2574
[139] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M,

Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9. https://doi.org/10.1038/s41586-021-03819-2

[140] Freed AS, Garde S, Cramer SM. Molecular simulations of multimodal ligand-protein binding: elucidation of binding sites and correlation with experiments. J Phys Chem B 2011;115:13320–7. https://doi.org/10.1021/Jp2038015

[141] Al-Thelaya K, Agus M, Gilal NU, Yang Y, Pintore G, Gobbetti E, Calí C, Magistretti PJ, Mifsud W, Schneider J. InShaDe: invariant shape descriptors for visual 2D and 3D cellular and nuclear shape analysis and classification. Comput Graph 2021;98:105–25. https://doi.org/10.1016/j.cag.2021.04.037

[142] Li C, Ben Hamza A. Spatially aggregating spectral descriptors for nonrigid 3D shape retrieval: a comparative survey. Multimed Syst 2014;20:253–81. https://doi.org/10.1007/s00530-013-0318-0