Received: 30 September 2021 | Revised: 23 December 2021 | Accepted: 19 January 2022

DOI: 10.1002/wcms.1602

FOCUS ARTICLE

WIREs COMPUTATIONAL MOLECULAR SCIENCE WILEY

Using machine-learning-driven approaches to boost hot-spot's knowledge

Nícia Rosário-Ferreira[1,2] | **Alexandre M. J. J. Bonvin**[3] | **Irina S. Moreira**[4,5]

[1]CQC—Coimbra Chemistry Center, Chemistry Department, Faculty of Science and Technology, University of Coimbra, Coimbra, Portugal

[2]CIBB—Center for Innovative Biomedicine and Biotechnology, University of Coimbra, Coimbra, Portugal

[3]Bijvoet Centre for Biomolecular Research, Science Faculty/Chemistry, Utrecht University, Utrecht, The Netherlands

[4]Department of Life Sciences, University of Coimbra, Coimbra, Portugal

[5]CNC - Center for Neuroscience and Cell Biology/CIBB, University of Coimbra, Coimbra, Portugal

Correspondence
Irina S. Moreira, Department of Life Sciences, University of Coimbra, Calçada Martim de Freitas, 3000-456 Coimbra, Portugal; CNC—Center for Neuroscience and Cell Biology, CIBB—Center for Innovative Biomedicine and Biotechnology, University of Coimbra, 3004-535 Coimbra, Portugal.
Email: irina.moreira@cnc.uc.pt

Funding information
European Union Horizon 2020 projects BioExcel, Grant/Award Numbers: 675728, 823830; COMPETE 2020-Operational Programme for Competitiveness and Internationalization and Portuguese National Funds via Fundação para a Ciência e a Tecnologia, Grant/Award Numbers: DSAIPA/DS/0118/2020, LA/P/0058/2020, POCI-01-0145-FEDER-031356, UIDB/04539/2020, UIDP/04539/2020

Edited by: Peter R. Schreiner, Editor-in-Chief

Abstract

Understanding protein–protein interactions (PPIs) is fundamental to describe and to characterize the formation of biomolecular assemblies, and to establish the energetic principles underlying biological networks. One key aspect of these interfaces is the existence and prevalence of hot-spots (HS) residues that, upon mutation to alanine, negatively impact the formation of such protein–protein complexes. HS have been widely considered in research, both in case studies and in a few large-scale predictive approaches. This review aims to present the current knowledge on PPIs, providing a detailed understanding of the microspecifications of the residues involved in those interactions and the characteristics of those defined as HS through a thorough assessment of related field-specific methodologies. We explore recent accurate artificial intelligence-based techniques, which are progressively replacing well-established classical energy-based methodologies.

This article is categorized under:

Data Science > Databases and Expert Systems

Structure and Mechanism > Computational Biochemistry and Biophysics

Molecular and Statistical Mechanics > Molecular Interactions

KEYWORDS

binding hot-spots, computational alanine scanning mutagenesis, interaction energetics, machine-learning algorithms, protein–protein interactions

1 | INTRODUCTION

Protein–protein interactions (PPIs) are major components of cellular communication. They form a complex and sophisticated network known as the "interactome." It is estimated that the whole human interactome consists of 130 k–650 k binary PPIs.[1,2] The interactome has a fundamental role in physiological and pathological processes such as cell growth, cell differentiation, apoptosis, signal transduction, and immune response.[3] Aberrant regulation of these protein–protein networks is known to be associated with many diseases including cancer, neurodegenerative, and infectious diseases, among others.[4,5] Recent studies showed that single amino acid variations (SAVs) typically found on PPI sites[5] can interfere (positively or negatively) with protein stability and/or complex formation,[6,7] hence affecting the downwards reaction/communication cascade.

Given their ubiquitous existence and involvement in disease, PPIs have been receiving increased attention as therapeutic targets. PPI targeting can potentially avoid the promiscuous effects on interactions pathways intrinsic to many existing drugs, especially when targeting a hub-protein.[8] Despite intensive efforts, developing PPI modulators remains a rather daunting challenge for many reasons. First, the average interface area of PPIs ranges from 1500 to 3000 Å^2, much larger than the average contact area of small molecule binding pockets (300–1000 Å^2). This results in a high-affinity binding between interacting proteins, increasing the design complexity of small-compounds targeting those interactions. Second, most protein–protein interfaces have topographically shallower surfaces than the lock-and-key-like deep grooves and pockets found for conventional drug targets.[9] Third, noncontiguous binding regions can occur, depending on the size of the interacting partners. Fourth, we are still missing a detailed understanding of the contribution of flexibility, dynamics, (partial) folding events, and such[10] to the establishment of correct PPIs; transient pockets might be involved in this process.[11] Fifth, the absence of small molecule endogenous ligands for PPIs as starting points constitutes an enormous test for structure-based drug design.[12–14] Despite the number of available PPI modulators being relatively small, PPIs are no longer considered uniformly undruggable due to the rapid advances of structural biology and related methodologies.[14] In fact, their modulation constitutes an important strategy undertaken by a variety of pharmaceutical industries and research groups,[14] and a variety of PPI-directed drugs are now approved, especially in the oncology area.[15]

Size and shape of protein–protein interfaces varies a lot. It is well known that protein regions have distinctive abilities to interact with other proteins, nucleic acids, and ligands, depending on their local curvature and physiochemical composition.[16] Pioneer work by Wells et al.[17–19] showed that only a small number of residues are truly responsible for the binding free energy, the so-called hot-spots (HS), which constitute entry points for PPI modulators design.[20] Besides small-molecules, PPI modulators can also be antibodies or peptides, depending on a variety of factors including interface size and polarity.[14] PPIs and HS are essential for a variety of systems, including, for example, viral infection, which is why researchers working to find new drugs to end this corona virus disease 2019 (COVID-19) pandemic are giving it special attention.[21–29]

In this review, we focus first on the main physico-chemical and structural characteristics of protein–protein interfaces and HS (Sections 2 and 3). In Section 4 we then present existing HS prediction algorithms for different interface types, with special emphasis on protein–protein complexes and artificial intelligence (AI)-based methodologies. A critical comparison is made regarding the used features and results achieved.

2 | PROTEIN-BASED INTERFACES

Various structural and sequence analyses of PPIs have been performed in previous years as these are fundamental to better understand the potential of key residues as HS, which will help find new therapeutic options.[30–33] One of the earliest classifications of protein–protein complexes was based on their lifetime and divided them as permanent (complex only stable in its oligomeric form) or transient (associates and dissociates *in vivo*). The distinction is sometimes difficult due to the overcrowding in the cytoplasm.[34] Some specific knowledge about the main characteristics of these types of complexes could already be recognized. Generally, permanent interfaces display higher co-expression and conservation rates than transient ones.[35] On the other hand, interfaces in transient interactions are smaller in size,[36] with a composition profile much more like the general surface and higher number of polar residues than permanent interfaces.[37] Jayashree et al. also showed that over 75% of the amino acids at protein–protein transient interfaces are involved in bifurcated interactions, where residues take part in both interprotein and intraprotein interactions simultaneously. It was also postulated that the microenvironment around these residues is preformed and maintained after complex

formation.[38] Transient interactions are also often mediated by disordered protein segments, small linear motifs and can require posttranslational modifications (PTMs),[37,39] complicating even further the understanding of their binding mechanisms.[10] Additionally, PTMs at functional sites can create new binding sites within transient pockets.[40] As such, transient complex formation is much harder to experimentally characterize.[33,41]

Another classification of protein–protein complexes distinguishes between obligate, if the monomeric forms are nonfunctional or unstable on their own *in vivo*, and nonobligate, if the monomers are stable and can exist on their own.[42,43] The interfaces of obligate complexes are usually larger and enriched in hydrophobic and aromatic residues whereas nonobligate interfaces are smaller and more polar.[43] Still, some proteins can change from nonobligate to obligate forms, depending on the cellular conditions.[43] Residues in protein–protein complexes can be split into:

1. Protein core, in which all residues are occluded from the solvent;
2. Protein surface, in which residues that have a relative surface area above a 0.20-0.25 cut-off value[44–46] (the relative surface area is defined as the ratio between the measured solvent accessible surface area (SASA) of a residue X and its corresponding area in a Gly-X-Gly peptide–$_r$SASA; and
3. Protein interface, corresponding to surface residues with one of their atoms within a 5 Å distance of any atom of a residue from the binding partner.

Protein–protein interfaces are well packed regions with a high degree of chemical and physical complementarity and with an inherent plasticity.[47] These regions are not rigid, with most of the flexibility coming from loop perturbation besides the classic sidechain movement.[48] It was also shown that only 26% of all interfacial residues exist in an α-helix, 24% within a β-strand, whereas the remaining ones do not possess a regular secondary structure.[48]

Although the diversity is high, PPIs could be split by interface size into three categories: (i) small, 1150–1200 Å^2, (ii) standard-size, 1200–2000 Å^2; and (iii) large, 2000–4660 Å^2.[49] Smaller protein complexes share physical elements common to the more traditional enzyme targets concept of lock-and-key: (i) high affinity within a relatively small surface area; and (ii) deeper pockets engaged by less than five major contributing amino acids to the binding free energy.[50] Besides contact area, binding affinity can also be used to classify protein–protein interfaces, further splitting them into four classes based on whenever they are narrow (surface area <2500 Å^2) or wide (surface area >2500 Å^2), and tight (K_d <200 nM) or loose (K_d >200 nM). Among them, the "narrow and tight" PPIs are more amenable to the design of small-molecule inhibitors.[50,51] Protein–protein interfaces consist of complex, uneven areas that involve the surface amino acids of a protein. Interfacial residues were further characterized using a three-layer model: (i) core (Equation (1)), buried residues in the interface with higher hydrophobicity and conservation, and small mobility; (ii) rim (Equation (2)), partially buried, flexible interface residues; and (iii) support (Equation (3)), amino acids with a composition that resembles the buried interior of a protein.[46,52,53]

$$\text{Core} = \Delta_r\text{SASA} > 0 \,\&\, {}_r\text{SASA}_m > 25\% \,\&\, {}_r\text{SASA}_c < 25\% \quad (1)$$

$$\text{Rim} = \Delta_r\text{SASA} > 0 \,\&\, {}_r\text{SASA}_m > 25\% \quad (2)$$

$$\text{Support} = \Delta_r\text{SASA} > 0 \,\&\, {}_r\text{SASA}_m < 25\% \quad (3)$$

where Δ_rSASA is the difference in relative solvent accessible surface area between monomer and complex ($_r$SASA$_m$–$_r$SASA$_c$) and $_r$SASA$_m$ and $_r$SASA$_c$ are the relative SASA in the monomer and in the complex, respectively. Single amino acids variations are more prone to occur at the interface core, with a propensity 2.1 higher than for the rim region, as the interface core is more conserved and more enriched in binding HS.[54,55]

Interfaces between proteins and nucleic acids, either protein–deoxyribonucleic acid (protein–DNA) or protein–ribonucleic acid (protein–RNA), exhibit a few differences compared to protein–protein interfaces. In particular, it seems that intrinsic conformational flexibility is particularly relevant for protein and nucleic acid complexes,[56] which are known to adapt their conformation to their binding partner.[57] Besides being more flexible, the interface residues are often more conserved, particularly at backbone-contacting positions.[58] Protein–DNA interfaces are nonobligate as both molecules exist in isolation as well as in the complex.[59] They involve on average 24 mainly positive and polar residues and 12 nucleotides.[60] It is further acknowledged that Arg is, by far, the most common amino acid at protein–nucleic acids interfaces as its side chain can establish multiple hydrogen bonds with the DNA phosphate, sugar, or nucleobase

moieties.[61] The second most enriched amino acid is Lys, followed by His, Ser, and Thr for DNA and Asn, His, and Gln for RNA complexes.[62] Tyr is also favored in both. The presence of these polar and charged residues shows that electrostatic is indeed the main driving force in these complexes by establishing multiple hydrogen-bonds, often water-mediated.[63] Hydrogen bonds established between amino acid sidechains and nucleotides bases seem to significantly contribute towards specificity, whereas the ones established with the phosphate backbone are more relevant for stabilization and orientation of the complex.[64] Protein–DNA and protein–RNA interactions show, however, differences as in the first, most hydrogen bonds involve phosphate atoms and in the second, base edge and ribose atoms.[62] Recent studies show the rich diversity of hydrogen-bonding interactions at these interfaces, highlighting their role for protein–nucleic acid recognition.[65] While hydrogen bonds account for nearly 50% of all interactions, van der Waals (vdW), and hydrophobic interactions are also common in these complexes.[62] Corsi et al., inspired by the support–core–rim model defined for protein–protein interfaces, defined a three archetypal model of protein–DNA interfaces, namely seed–extension–outer layer.[66] They used evolutionary conservation, physico-chemical properties, and local/global geometry to classify residues for their role in the protein–DNA interaction.[66]

3 | BINDING HOT-SPOTS

Wells *et al.* first applied systematic experimental alanine scanning mutagenesis (ASM) to probe PPIs in a complex between human growth hormone and its receptor. They defined HS as those residues that resulted in at least a 2.0 kcal/mol difference in the binding free energy between mutant and wild type ($\Delta\Delta G_{\text{binding}}$, Equation (4)) upon point alanine mutation.[17]

$$\Delta\Delta G_{\text{binding}} = \Delta G_{\text{mutant}} - \Delta G_{\text{wild-type}} \tag{4}$$

The degree of evolutionary conservation of an amino acid residue in a protein reflects a balance between its natural tendency to mutate and its importance in the preservation of structural integrity and/or function of the protein. It was postulated that HS are usually conserved amino acid positions, evolving more slowly, as they are essential to maintain the proper binding mode of a protein complex, and thus its function. Amino acids have different propensities to be HS; Arg, Tyr, and Trp are the most frequent ones due to their conformation, size and potential to establish meaningful interactions such as hydrophobic contacts, hydrogen bonds, electrostatic interactions, and π–π stacking.[20] SAVs[55] were found to be highly associated with these residues (Arg > Trp > Tyr > Gln > His > Gly > Cys).[54] For example, Arg accounts for 16%–19% of disease-causing SAVs, predominantly if located at the interface core regions.[55,67]

HS were found to be clustered and packed to form "hot regions"[68] in complemented pockets, and are disfavored in unfilled pockets, the ones that remain empty after protein–protein complexation.[69] These regions were also assessed by alanine shaving, the concerted mutation of two or more interfacial residues, to evaluate cooperativity.[70] For example, Moreira et al. showed that aromatic HS residues are especially relevant for protein–protein complexes and enriched near other HS to form π–π and cation–π interactions within cooperative high-order clusters.[71] It was indeed proposed that HS contributions to the binding energy is additive between "hot regions" whereas cooperative within a "hot region," maybe due to local or global changes of the protein conformation, solvent structure, or other protein dynamic properties.[68,70] However, other studies point to long-range cooperative effects.[72] Kuttner et al. demonstrated that the backbone dynamic landscapes of these interacting surfaces form "stability patches" for which a diminished enthalpy–entropy compensation effect is key.[73,74]

The formation of these "hot regions" implies that HS from opposite monomers face each other and are generally enriched at the center of the binding protein–protein interface. Bogan and Thorn showed that these regions are typically surrounded by energetically fewer essential residues, resembling an O-ring, whose function seems to occlude HS from bulk water molecules.[20,75–77] This "O-ring theory" or "Water Exclusion" hypothesis implies that HS exist within a low dielectric environment with a low solvent exposure, favoring the establishment of relevant interactions.[20,75–77] As such, most computational HS detection methods use an energy term/feature related to solvation.[78] Ramos et al. also demonstrated that HS in protein–DNA complexes tend to be occluded from the solvent, extending the applicability of the O-ring theory to other protein-based complexes.[77]

Even though HS are main contributors for binding affinity and stability, not all are fundamental for specificity, as they could be shared among different partners, particularly if within a hub-protein. In fact, hub-proteins can be split into "date" and "party" hubs, whenever the interactions occur discretely, using the same or overlapping interfaces, or

simultaneously, using multiple interfaces on its surface to couple to various partners.[79] "Date" hubs reutilize HS in different ways to perform different functions.[80] In an analogy to HS, Gavenonis et al. used the term hot-loops (HL) to classify a set of loops (5.6% of the overall interface) that significantly contribute to binding interactions.[81] They revealed that 36% of HL were responsible for more than half of all interfaces binding energy. The typical residues found on these regions were Trp, Phe, His, Asp, Tyr, Leu, Glu, Ile, and Val.[81] Camacho et al. have also introduced the term "anchor residues." These residues were defined as the ones with ΔSASA higher than 0.5 Å^2 upon complex formation and a binding free energy difference higher than 0.5 kcal/mol.[82] Cold-spots (CS), another concept introduced by Shirian et al., are residues where three or more different substitutions lead to at least 0.3 kcal/mol improvement in binding affinity (decrease in $\Delta\Delta G_{binding}$—Equation (4)).[80] They showed that CS were positions at the wild-type complex where the intermolecular interactions were not optimal.

HS have also been shown to correspond to key binding regions, able to couple small molecule ligands.[17,83] Concurrent with the development of protein–protein HS, protein binding sites are also explored to detect druggable HS using fragment size or organic probes. Zerbe et al. showed that this HS concept is largely complementary with PPI ones with a few additional topological requirements.[84] These druggable PPI regions were shown to have a higher number of aromatic residues and methionines.[85] FTMap is a well-known consensus strategy that uses organic probes within a grid to identify binding HS and new binding sites to small molecules.[86–88] Kozakov et al. showed that only fragments with a good spatial overlap with top-ranked HS were expected to be extended to larger, useful ligands.[89] However, when dealing with a shallower protein–protein interface, the lack of protein flexibility may introduce demanding problems to the detection of key HS.[90] Molecular dynamics (MD) application, although very useful to overcome this issue, continues to be time consuming, and as such the systematically use of organic/aqueous mixed solvents has been proposed to predict binding modes and affinities, or to guide the fragment evolution process. One example was the recent development of FragMaps.[91] Energetics and plasticity were also assessed by Mertz et al. in their binding HS identification algorithm, used to predict ligand binding modes.[92] More recently, Bajusz et al. developed SpotXplorer0 library, a minimal set of fragment pharmacophores upon critical analysis of HS at target proteins.[93] More details about these methods can be found in Table 1.

As protein–protein surfaces are not rigid, their inherent conformational fluctuation can open pharmacologically relevant transient pockets that are important for the binding of new drugs.[164–166] Indeed, the knowledge of such druggable HS has been shown to help identifying transient pockets in interleukin-2 complexes.[166,167] Moreover, transient PPIs (TPPIs)[33] are also involved in a variety of disease-related pathways, and a few drugs were found to bind via "interfacial inhibition."[168] This mechanism focuses on the drug binding to transient exposed HS at a protein–protein complex, stabilizing its normally transient transition state, a structurally and energetically unbalanced state.[169] A few in silico methods were already developed to identify these cryptic pockets or to better characterize TPPIs, and typically involve MD simulations to surpass the lack of experimental structures and facilitate in-depth analysis of structural, functional dynamic aspects of PPI models.[170] For example, Rosell et al. used a combination of MD-generated side-chain conformers, which produced thousands of transient cavities across the protein surface, and protein–protein docking methods to find druggable HS.[171]

4 | IN SILICO METHODOLOGIES FOR HS IDENTIFICATION/PREDICTION

4.1 | Databases

Experimental ASM involves the systemic point mutation of binding interface positions, followed by expression and purification of mutants and measurement of their binding affinities. These experiments are time consuming and labor intensive, highly depend on the used assays, and consequently not widely applied. A few databases with available experimental information are listed in Table 2, some of which gather information from other mutagenesis experiences besides alanine. For protein–protein complexes there are four main databases: the alanine scanning energetics database (ASEdb),[11] protein–protein complex mutation thermodynamics (PROXiMATE,[174] previously known as PINT[172]), the binding interface database (BID[116]), and structural database of kinetics and energetics of mutant protein interactions (SKEMPI), whereas for protein–nucleic acid, we can access protein–nucleic acid interactions (PRONIT[176]) and protein–nucleic acid binding energetic database (NABE[177]). Table 2 also includes some other curated, nonredundant datasets of mutations that satisfy a few requirements:

TABLE 1 Computational HS detection methods available in the literature

Type	Acronym	Method name	Used CPX	Used MUTS	Used HS	Algorithm	Features	Evaluation	Methodologies	Year	References
PL	FTMap	Fourier transform mapping	2 (6600 cpx after Step 1)	NA	NA	FFT Simple greedy algorithm	Energy function	Qualitative evaluation	AI-structure- and energy-based	2002	88,94,95
PL	SpotXplorer	NA	NA	NA	NA	NA	Fragment and pharmacophore approach using FTMap	NA	AI-structure- and energy-based	2021	93
PL	HS-Pharm	Hot-spots (HS)-guided receptor-based pharmacophores	3500 binding cavities	NA	NA	RF J48 decision trees Naive Bayesian inference	26 attributes for CFP1 and CFP2 fingerprints and 52 attributes for CFP3 fingerprint 4 topological and physico-chemical properties of protein cavity atoms	Various metrics. Please check publication	AI-structure-based	2008	96
PL	FTFlex	Flexible protein mapping	15	NA	NA	RMSD, correlation coefficient	NA	NA	AI-structure-based	2013	97
PL	NA	Furtmann et al.	580	NA	630 (3D-cliffs)	NA	NA	NA	AI-structure-based	2015	98
PL	NA	Mertz et al.	1	NA	NA	MD simulation MM/PB(GB)SA	NA	NA	Energy-based	2012	92
PNA	PreHots	Predicting hot spots	89	NA	123	Sequential backward method Catboost, XGBoost GTB Logistic regression classifier	19 Network, exposure, sequence, and structure features	F1: 0.74 ACC: 0.77	AI-sequence- and structure-based	2020	99
PNA	NA	Munteanu et al.	28	177	*43*	SVM RF Bayesnet	Conservation, SASA related features	F1: 0.71 AUC: 0.65	AI-sequence- and structure-based	2015	100
PNA	PrabHot	Prediction of protein–RNA binding hot spots	47	NA	107	GTB SVM ERT	125 network, exposure, sequence, and structure features	F1: 0.75 AUC: 0.86	AI-sequence- and structure-based	2018	101
PNA	XGBPRH	NA	47	209	107	McTWO XGBoost	6 features (2 network, 2 exposure, and 2 structural features)	F1: 0.87 AUC: 0.87	AI-Sequence- and Structure-based	2019	102
PNA	iPNHOT	Identification of protein–nucleic acid interaction HOT spots	*Training dataset* 105 *Independent test dataset* 32	NA	*Training dataset* 86 *Independent test dataset* 14	SVM	SASA, conservation, sequence physiochemical	F1: 0.39 ACC: 0.61	AI-sequence- and structure-based	2020	103
PNA	sxPDH	S-ISOMAP and XGBoost based model for prediction of protein-DNA binding hot spots	64	NA	88	Supervised isometric feature mapping XGBoost	114 features from a combination of the protein sequence, structure, network, and solvent accessible information	F1: 0.71 AUC: 0.77	AI-sequence- and structure-based	2020	104

TABLE 1 (Continued)

Type	Acronym	Method name	Used CPX	Used MUTS	Used HS	Algorithm	Features	Evaluation	Methodologies	Year	References
PNA	SPHot	Sequence-based Prediction of Hot spots	47	NA	107	EVC oc RBF-based SVM Sigmoid-based SVM k-nearest neighbor	43 final predictors	F1: 0.84 AUC: 0.89	AI-sequence-based	2019	105
PNA	inpPDH	Interfacial neighbor properties protein DNA Hotspot	64	NA	88	SVM	7 hybrid features of traditional and new interfacial neighbor properties	F1: 0.731 AUC: 0.83	AI-structure-based	2021	106
PP	PCRPi	Presaging critical residues in protein interfaces	25	NA	78	Bayesian networks	Energetic, structure-based, and sequence-based features	F1: 0.71 ACC: 0.75	AI-sequence and structure-based	2009	107
PP PNA	CCRXP	Clusters of Conserved Residues	NA	150 mutations in PP interfaces 157 single-residue mutations in RNA-binding protein	58 hot-spots mutations 60 hot-spots mutations	NA	NA	NA	AI-sequence and structure-based	2010	108
PP	SpotON	Hot SPOTs ON protein complexes	53	534	127	svmPoly	881 features	F1: 0.96 AUC: 0.91	AI-sequence and structure-based	2017	109
PP	sbSVM	semi-supervised boosting SVM	*Dataset 1* 17 *Dataset 2* 17 *Independent test dataset* 18	NA	*Dataset 1* 65 *Dataset 2* 65 *Independent test dataset* 39	SemiBoost framework SVM with semi-supervised boosting	Top 10 features after RF of 6 sequence features and 62 structure features	*Independent test dataset (trained on Dataset 1)* F1: 0.58 ACC: 0.66 *Independent test dataset (trained on Dataset 2)* F1: 0.63 ACC: 0.70	AI-sequence- and structure-based	2012	110
PP	NA	Munteanu et al.	*Dataset 1* 15 *Dataset 2* 15 *Dataset 3* 28	*Dataset 1* 477 *Dataset 2* *91* *Dataset 3* 222	*Dataset 1* 80 *Dataset 2* *35* *Dataset 3* *79*	SVM RF Bayesnet	Conservation, SASA related features	*Dataset 1* F1: 0.83 AUC: 0.85 *Dataset 2* F1: 0.68 AUC: 0.75 *Dataset 3* F1: 0.65 AUC: 0.62	AI-sequence- and structure-based	2015	100
PP	HEP	NA	*Training set* 17 *Test set* 18	*Training set* NA *Test set* 127	*Training set* 62 *Test set* 39	SVM	Top 3 features from 108 sequence, structural and neighborhood features via mRMR	F1: 0.70 ACC: 0.79	AI-sequence- and structure-based	2016	111
PP	SBHD2	Sasa-based hot-spot detection 2	53	545	140	27 algorithms	38 structural and 41 genomic features	*Testing dataset* F1: 0.62 AUC: 0.69	AI-sequence- and structure-based	2016	112

(Continues)

TABLE 1 (Continued)

Type	Acronym	Method name	Used CPX	Used MUTS	Used HS	Algorithm	Features	Evaluation	Methodologies	Year	References
PP	RBHS	Robust principal component analysis-based prediction of PPI hot spots	*HN-34* 34 and *BID-18* 18	*HN-34* 313 and *BID-18* 126	*HN-34* 133 and *BID-18* 39	Principal component pursuit SciKit XGBoost	6 physico-chemical features, 5 solvent accessible area, 7 solvent exposure, 20 PSSM profiles, 20 block substitution matrices	F1: 0.66 ACC: 0.77	AI-sequence- and structure-based	2020	113
PP	NA	Chen et al.	NA	149 (ASEdb) 112 (BID)	58 (ASEdb) 54 (BID) 196 (SKEMPI)	IBL	132 physiochemical features (AAindex1)	F1: 0.76	AI-sequence-based	2013	114
PP	NA	Hu et al.	ASEdb[115] BID[116] SKEMPI[117]	NA	235	IBL	33 top physico-chemical classifiers	F1: 0.77 (ASEdb) 0.80 (BID) 0.65 (SKEMPI)	AI-sequence-based	2017	118
PP	SPOTONE	Hot spots ON protein complexes with Extremely randomized trees	53	NA	127	ERT	173 sequence features	F1: 0.85 ACC: 0.82	AI-sequence-based	2020	119
PP	NA	Li et al.	15	296	83	Interaction's evaluation	Types of contacts	Successful rate of 0.71	AI-structure-based	2006	120
PP	K-CON	Knowledge-based biochemical contact analysis	*Independent test set* 19	*Independent test set* 112	*Independent test set* 50	Decision trees FADE	Biochemical contact features (shape-related features, atomic contacts, hydrogen bonds, salt bridges, chemical type)	F1: 0.48	AI-structure-based	2007	121
PP	K-FADE	Knowledge-based fast atomic density evaluation	*Independent test set* 19	*Independent test set* 112	*Independent test set* 50	Decision trees FADE	Shape specificity features (shape specificity, FADE points, residue size)	F1: 0.41	AI-structure-based	2007	122
PP	KFC	Knowledge-based FADE and contacts	*Cross validation set* 16 *Independent test set* 19	*Cross validation set* 249 *Independent test set* 112	*Cross validation set* 60 *Independent test set* 50	Combination of K-FADE and K-CON	Combination of K-FADE and K-CON	*Independent test set* F1: 0.42	AI-structure-based	2008	122,123
PP	NA	Grosdidier et al.	*Dataset 1* 21 *Dataset 2* 22	*Dataset 1* 586 *Dataset 2* 361	*Dataset 1* 168 *Dataset 2* 94	Docking	NA	*Dataset 1* PPV: 0.78 TPR: 0.24 *Dataset 2* PPV: 0.78 TPR: 0.15	AI-structure-based	2008	124
PP	NA	HotSprint	34,817	NA	NA	Rate4Site	NA	ACC: 0.76	AI-structure-based	2008	125

TABLE 1 (Continued)

Type	Acronym	Method name	Used CPX	Used MUTS	Used HS	Algorithm	Features	Evaluation	Methodologies	Year	References
PP	MINERVA	MINE residue VAlue	*T2 Dataset* 17 *Test set* 18	*T2 Dataset* 265 *Test set* 127	*T2 Dataset* 65 *Test set* 39	Decision tree SVM	18 chosen from 54 structural, sequence, and molecular interaction features (for T2 training set)	*Independent test set* F1: 0.57	AI-structure-based	2009	126
PP	NA	Demerdash et al.	*Training dataset* 11 *Test dataset* 5	NA	44	SVM	*Feature set 1* 5 dynamical, 9 structural, 2 network, and 2 informatic measures *Feature set 2* 21 structural measures defined by Daily and Gray	*Feature set 1* PPV: 0.58–0.67 TPR: 0.67–0.81 *Feature set 2* PPV: 0.55–0.59 TPR: 0.81–0.92 *Feature set combination* PPV: 0.73–0.81 TPR: 0.64–0.71	AI-structure-based	2009	127
PP	NA	Guharoy et al.	13	462	NA	$\Delta\Delta G$ ΔASA	NA	Fraction correctly predicted (total, core) of 0.82, 0.83	AI-structure-based	2009	128
PP	NA	Higa et al.	Same as Refs. 122,129	Same as Refs. 122,129	Same as Refs. 122,129	SVM	43 structural and evolutionary parameters	F1: 0.60	AI-structure-based	2009	130
PP	APIS	A combined model based on protrusion index and solvent accessibility	*Training set* 17 *Independent test set* 18	*Training set* NA *Independent test set* 127	*Training set* 62 *Independent test set* 39	SVM	9 structural features	F1: 0.64	AI-structure-based	2010	131
PP	GCR	Geometrically centered region	13	355	109	Atom burial level determined by Dijkstra's algorithm	Voronoi diagram	F1: 0.58	AI-structure-based	2010	132
PP	HotPoint	NA	NA	*Training set* 150 *Test set* 112	*Training set* 58 *Test set* 54	Decision tree	Accessibility, conservation, pair potentials, computational alanine scanning	F1: 0.65 ACC: 0.70	AI-structure-based	2010	133,134
PP	DBAC	Deeply buried atomic contacts	13	258	50	SVM	Feature set of the deeply buried atomic contacts only	F1: 0.62 ACC: 0.86	AI-structure-based	2011	135
PP	NA	Chen et al.	28	904	107	RS-MCLP	9 relevant features for HS prediction from the 14 evaluated structural and physiochemical features	F1: 0.28	AI-structure-based	2011	136

(Continues)

TABLE 1 (Continued)

Type	Acronym	Method name	Used CPX	Used MUTS	Used HS	Algorithm	Features	Evaluation	Methodologies	Year	References
PP	SemiHS	NA	17	265	65	Iterative semi-supervised SVM	8 sequence, 5 structure and 4 energy features	AUC: 0.85	AI-structure-based	2011	137
PP	KFC2	Knowledge-based FADE and contacts	*Cross validation set* 17 *Independent test set* 18	*Cross validation set* 265 *Independent test set* 126	*Cross validation set* 65 *Independent test set* 39	SVM	47 initial interface solvation, atomic density, and plasticity features KFC2a has 8 features that are mainly related to SASA and local plasticity KFC2b has 7 other features (5 exclusive)	*KFC2a (Independent set)* ACC: 0.75 *KFC2b (Independent set)* ACC: 0.78	AI-structure-based	2011	138
PP	NA	Wang et al.	*Training set* 20 *Test set* 18	*Training set* 318 *Test set* 125	*Training set* 79 *Test set* 39	RF with hybrid features	5 category descriptors (23 physiochemical features)	*Alanine scanning database* ACC: 0.82 *BID* ACC: 0.78	AI-structure-based	2012	139
PP	NA	Ozbek et al.	33	NA	173	GNM	Structural features	*Unbound structures* ACC: 0.81–0.92 *Complex structures* ACC: 0.94–0.97	AI-structure-based	2013	140
PP	SBHD	Sasa-based hot-spot detection	15	248	65	SVM	SASA-related features	F1: 0.86 ACC: 0.77	AI-structure-based	2013	141
PP	NA	Wang et al.	*Training set* 20 *Independent test set* 18	*Training set* 318 *Independent test set* 125	*Training set* 77 *Independent test set* 38	Random forest (RF)	Hybrid features with target and spatial neighbor residues information	ACC: 0.77	AI-structure-based	2014	142
PP	NA	Hot loops	1,242 from 9,388	Alanine scanning	NA	NA	NA	NA	AI-structure-based	2014	81
PP	PredHS	prediction of hot spots	*Dataset 1* 17 *Dataset 2* 17 *Independent test set* 18	*Dataset 1* 265 *Dataset 2* NA *Independent test set* 127	*Dataset 1* 65 *Dataset 2* 65 *Independent test set* 39	*Feature selection* RF and sequential backward elimination method SVM PredHS-Ensemble	38 structural neighborhood properties	*Independent dataset* *PredHS-SVM* F1: 0.68 ACC: 0.83 *PredHS-Ensemble* F1: 0.68 ACC: 0.79	AI-structure-based	2014	143,144
PP	βACV_{ASA}	β contact's atomic contact vector accessible surface area	22	396	86	βACV_{ASA} Ridge regression	An ACV of β contacts with ASA integration	F1: 0.60 ACC: 0.83	AI-structure-based	2014	145
PP	NA	Li et al.	20	471	86 ($\Delta\Delta G > 2.0$ kcal/mol) 180 ($\Delta\Delta G > 1.0$ kcal/mol)	Onion-like model (atomic contact graphs and burial level patterns)	Δ Burial level	$\Delta\Delta G > 1.0$ kcal/mol F1: 0.68 $\Delta\Delta G > 2.0$ kcal/mol F1: 0.46	AI-structure-based	2015	146

TABLE 1 (Continued)

Type	Acronym	Method name	Used CPX	Used MUTS	Used HS	Algorithm	Features	Evaluation	Methodologies	Year	References
PP	ppRF	NA	*ASEdb* 20 and *Independent test set* 36	*ASEdb* 366 and *Independent test set* 232	*ASEdb* 79 and *Independent test set* 53	RF	4 types of features encompassing 143 individual features associated with each ΔΔG	F1: 0.57 ACC: 0.77	AI-structure-based	2015	147
PP	NA	Zhang et al.	32	NA	171	GNM GNB	*Feature set*: Top 20 highest frequency modes *2 coding schemes about the feature vectors*: Varying distance cutoffs for GNM and sliding window sizes for GNB	F1: 0.15 ACC: 0.80	AI-structure-based	2016	148
PP	iPPHOT	NA	*Training set* 15 *Independent test set* 16	*Training set* 154 *Independent test set* 95	*Training set* 62 *Independent test set* 28	SVM	6 sequence and structure features selected via decision tree and mRMR by using a PSFS	F1: 0.62 ACC: 0.71	AI-structure-based	2018	149
PP	NA	Li et al.	*Dataset 1* 28 *Dataset 2* NA	*Dataset 1* 797 *Dataset 2* 155	*Dataset 1* 107 *Dataset 2* 65	mRMR SVM LCSD	8 physico-chemical features	*HS prediction* F1: 0.72 *Hot Regions prediction* F1: 0.81	AI-structure-based	2018	150
PP	PredHS2	Prediction of hot spots 2	34	313	NA	XGBoost	26 optimal features using mRMR and sequential forward selection process	F1: 0.79 ACC: 0.87	AI-structure-based	2018	151
PP	PIIMS	Protein interface in silico mutation scanning	50	1341	NA	MD simulation One-step FEP	NA	*Alanine mutations* ACC: 0.85	AI-structure-based	2021	152
PP	NA	Almlöf et al.	2	52	NA	MD simulation Optimized LIE	Free energy calculations	MUE: 0.50–1.03 kcal/mol	Energy-based	2006	153
PP	NA	Moreira et al.	3	47	11	MD Simulation MM-PBSA	Different dielectric constants	Mean and maximum error of 0.80 and 4.52 kcal/mol	Energy-based	2007	184
PP	NA	Lise et al.	20	349	81 ($\Delta\Delta G > 2.0$ kcal/mol) 165 ($\Delta\Delta G > 1.0$ kcal/mol)	Transductive SVM Gaussian processes	12 input features of energy components	F1: 0.60	Energy-based	2009	154

(Continues)

TABLE 1 (Continued)

Type	Acronym	Method name	Used CPX	Used MUTS	Used HS	Algorithm	Features	Evaluation	Methodologies	Year	References
PP	NA	Carbonell et al.	877	NA	76% residues predicted HS in specific binding sites and 24% in promiscuous binding sites	Agglomerative hierarchical algorithm, FoldX, edge-betweenness algorithm, RMSD, binding free energy	Specificity and affinity of PP interactions	NA	Energy-based	2009	155
PP	HSPred	HotSpot prediction	20 (+2 with experimental $\Delta\Delta G$ values)	349 (+16)	81 ($\Delta\Delta G > 2.0$ kcal/mol) 165 ($\Delta\Delta G > 1.0$ kcal/mol)	SVM	3 Classifiers (SVM_X, SVM_E, SVM_R) with 7, 3, and 4 input features of energy components	F1: 0.65	Energy-based	2011	156
PP	NA	Ramos et al.	3	46	28	MM/PB(GB)SA MD simulation	Internal dielectric constants	Overall success of 0.80 MUE: 0.80 kcal/mol	Energy-based	2013	157
PP	NA	Simões et al.	15	210	92	cASM MD simulation	Dielectric constants for nonpolar, polar, and charged residues	*cASM (set 7,7,11)* >1.5 kcal/mol ACC: 0.75 *Geometry optimization* MUE: 1.4 kcal/mol	Energy-based	2017	158
PP	HotSpot Wizard	NA	NA	NA	NA	NA	Webserver using consensus	NA	Energy-based	2018	159
PP	NA	Oshima et al.	18	341	70 ($\Delta\Delta G > 2.0$ kcal/mol) 146 ($\Delta\Delta G > 1.0$ kcal/mol)	Combines the water-entropy gain theory with the morphometric approach	NA	*$\Delta\Delta G > 2.0$ kcal/mol* F1: 0.545 *$\Delta\Delta G > 1.0$ kcal/mol* F1: 0.673	Energy-based	2011	160
PP	NA	Carl et al.	4	NA	7	ProBiS Web server CHARMM MD simulation	NA	NA	Energy-based	2012	161
PP	BudeAlaScan	Bristol university docking engine Alanine Scan	NA	748 mutations from SKEMPI	NA	ISAMBARD[162] Free energy calculations	Physico-chemical features	Pearson R of 0.50 Fractions correct of 0.76	AI-structure- and energy-based	2019	163

Note: Equations: F1-score $= \frac{2\times TP}{2\times TP+FP+FN}$; ACC $= \frac{TP+TN}{TP+FP+FN+TN}$; TPR $= \frac{TP}{TP+FN}$; PPV $= \frac{TP}{TP+FP}$.

Abbreviations: AI, artificial intelligence; AUC, area under the curve; cASM, computational alanine scanning mutagenesis; CFP, cavity fingerprint; CHARMM, chemistry at Harvard macromolecular mechanics; ERT, extremely randomized trees; F1, F1-score; FADE, fast atomic density evaluation; FDR, false discovery rate; FEP, free energy perturbation; FFT, Fourier fast transform; FN, false negatives; FNR, false negative rate; FP, false positives; GNB, gaussian naive bayes; GNM, gaussian network models; GTB, gradient tree boosting; IBL, instance-based learning; ISAMBARD, intelligent system for analysis, model building, and rational design; LCSD, local community structure detecting; LIE, linear interaction energy; MAE, mean absolute error; MCC, Matthew's correlation coefficient; MD, molecular dynamics; MM-PB(GB)SA, molecular mechanics Poisson–Boltzmann (generalized Born) surface area; MM-PBSA, molecular mechanics, Poisson–Boltzmann surface area; mRMR, maximum relevance–minimum redundancy; MSE, mean squared error; MUE, mean unsigned error; NA, not available; NIP, normalized interface propensity; NPV, negative predictive value; PCC, Pearson correlation coefficient; PL, protein–ligand; PNA, protein–nucleic acid; PP, protein–protein; PPV, positive predictive value/precision; RBF, radial basis function; RMSE, root mean squared error; RS-MCLP, rough set-based multiple criteria linear programming; SCC, Spearman's ρ; SR, success rate; SVM, support vector machine; TN, true negatives; TNR, true negative rate/specificity/selectivity; TP, true positive; TPR, true positive rate/sensitivity/recall; XGBoost, extreme gradient boosting.

1. An existing set of relative binding free energy difference ($\Delta\Delta G_{binding}$) values for interfacial residues coming from experimental alanine mutagenesis;
2. Availability of a three-dimensional (3D)-structure in the protein databank (PDB); and
3. A maximum of 35% sequence identity in each interface, hence preventing repeated complexes.

4.2 | Classical prediction methods

To overcome the problems inherent to experimental procedures, we have witnessed for the last two decades the raise of in silico methods for HS identification/prediction due to their lower cost, faster procedures, simplicity, and reliability[178] (Figure 1). Typically, algorithms for HS identification/prediction for protein–protein interfaces depend on the availability of a 3D structure, with a few exceptions: SpotONE,[119] based on features retrieved from protein sequence only such as one-hot encoding, relative position, amino acid basic knowledge and sliding window combinations of those, and HotSpotEC, an ensemble classifier based on SASA and physiochemical properties of amino acid sequences.[114,118] Table 1 also lists the available methods to detect HS on protein–protein and protein–nucleic acid systems, classifying those into atomistic, energy-based, or AI-based approaches, which vary on the type of used protein characteristics (structural and/or sequence-based).

Energy-based methods to perform computational ASM (cASM) have the advantage of providing quantitative analysis by capturing the free energy change upon alanine mutation and can be continuously improved either by the longer or multiple MD simulations and/or by using more accurate Hamiltonians (force fields). These approaches can be split into: (i) rigorous methods such as free energy perturbation (FEP)[179] and thermodynamic integration (TI)[180] or (ii) more simplistic approaches like molecular mechanics Poisson–Boltzmann (Generalized Born) surface area (MM/PB(GB)SA)[181–188] or other simple energy-based calculations. The MM/PB(GB)SA methodology combines a molecular mechanics approach with continuum solvent models for the calculation of the relative binding free energy (Equation (4)). For mutant and wild type, the binding free energy is the difference between the free energy of the complex and the two coupled monomers (Equation (5); e.g., protein A and protein B). The free energy of any involved molecule (Equation (6)) includes enthalpic and entropic contributions and is given by the sum of the internal covalent energies (bond, angles, and dihedrals), the electrostatic and the vdW nonbonded interactions, the polar solvation free energy, the nonpolar solvation free energy and the entropic contribution.

$$\Delta G_{\text{binding-molecule}} = G_{\text{complex}} - \left(G_{\text{protein\_A}} + G_{\text{protein\_B}}\right) \tag{5}$$

$$G_{\text{molecule}} = E_{\text{internal}} + E_{\text{electrostatic}} + E_{\text{vdW}} + G_{\text{polar\_solvation}} + G_{\text{nonpolar\_solvation}} - TS \tag{6}$$

For the calculations of the relative free energies between closely related complexes (point alanine mutant vs. wild-type), it is assumed that the total entropic term in Equation (6) is negligible as the partial contributions essentially cancel each other in Equation (4).[180,183,184,186,187,189] The $G_{\text{nonpolar solvation}}$ comes from the vdW interaction between the solute and the solvent, and it is proportional to the SASA value (Equation (7)).

$$G_{\text{nonpolar\_solvation}} = 0.00542 \times \text{SASA} + 0.92 \tag{7}$$

The $G_{\text{polar solvation}}$ can be more rigorously calculated by traditionally solving the linear Poisson–Boltzmann (LPB) equation or the nonlinear Poisson–Boltzmann (NLPB) equation, accounting for the importance of salt concentration in the medium (useful for protein–nucleic acid complexes). Poisson–Boltzmann is based on the second-order elliptic partial differential equation that describes the electrostatic potential surrounding a charge distribution. A variety of packages exist to solve this equation, such as Delphi[190] that uses a finite difference method, based on discretizing the workspace into a uniform grid. This continuum model involves a low dielectric protein surrounded by a high dielectric continuum solvent/water. Due to the elevate computational time involved, PB can also be substituted by an approximated method using the GB model.[191]

The MM/PB(GB)SA approach first developed by Massova *et al.*[187] was further improved by Moreira *et al.* that by using a set of three different internal dielectric constants (ε) to calculate $G_{\text{polar solvation}}$ that simulate the degree of

TABLE 2 Available datasets with experimental alanine mutagenesis data for protein-based systems with a known 3D complex structure

Type	Acronym	Name	Data	# HS	Webpage	Year	References
PP	NA	Combined from ASEdb,[115] BID,[116] SKEMPI, PINT[172] from Moreira et al.	534 nonredundant mutations from 53 complexes	127	https://alcazar.science.uu.nl/cgi/services/SPOTON/spoton/	2017	109
PP	NA	Alanine Scanning Energetics database (ASEdb)[115] and data from Kortemme et al.[173] from Cho et al.	265 mutants from 17 complexes	65	Not maintained anymore	2009	173
PP	BID	Binding interface database	126 mutants 18 complexes classified as "Strong," "Intermediate," "Weak," "Insignificant"	39 ("Strong")	Not maintained anymore	2003	116
PP	SKEMPI2.0	Structural database of kinetics and energetic of mutant protein interactions	2321 nonrepeated mutations from 180 complexes	358	https://life.bsc.es/pid/skempi2/	2019	117
PP	PROXiMATE—Previous version PINT—Protein–protein Interaction Thermodynamic[172]	PROtein–protein compleX MutAtion ThErmodynamics	1509 mutants from 89 complexes	263	http://www.iitm.ac.in/bioinfo/PROXiMATE/	2017	174
PNA	NA	Nonredundant dataset based on dbAMEPNI[175]—database of alanine mutagenic effects for protein–nucleic acid interactions from Zhu et al.	417 mutants from 137 complexes	100	http://zhulab.ahu.edu.cn/iPNHOT/ and http://zhulab.ahu.edu.cn/dbAMEPNI	2020	103
PNA	PRONIT	Database for PROtein–Nucleic acid InTeractions	177 mutants from 29 complexes	43	http://dna00.bio.kyutech.ac.jp/pronit/	2006	176
PNA	NABE	Protein–nucleic acid binding energetic database	1751 mutations from 405 complexes	240	http://nabe.denglab.org/	2021	177

Note: Cells in pink are also listed nonredundant datasets.
Abbreviations: NA, not available; PP, protein–protein; PNA, protein–nucleic-acid.

relaxation upon alanine mutation, achieving the chemical accuracy with a mean error of 0.80 kcal/mol.[184] Posteriorly, Martins *et al.* compared the method to TI for 22 mutants from four complexes and concluded that both presented similar accuracy (average error 1.18 vs. 1.53 kcal/mol, respectively), further validating the efficiency of the developed method.[180] MM/PB(GB)SA methods were also applied to protein–nucleic acid interfaces. For example, Ramos *et al.* following similar protocols also achieved a high accuracy.[185] More recently, other groups have been implementing a set of different ϵ to analyze the mutation effect at various types of interfaces.[192,193] These methods allow to access protein heterogeneous ensembles of fluctuating conformations and consider dynamics, flexibility, formation of transient interactions, and pockets. In contrast, as they rely on heavy MD simulations for conformational sampling at an atomistic resolution, typically in an explicit solvent representation, they are computational expensive and therefore difficult to apply in a high throughput mode.[9] Moreover, MM/PB(GB)SA still tend to neglect changes in the conformation entropy due to its large computational cost (S is neglected in Equation (6)). However, recent approaches have successfully used a new term, the interaction entropy (IE), combined with a MM/GBSA approach to calculate the binding free energy difference upon alanine mutation.[194]

4.3 | AI-based prediction methods

As energy-based methods are often time-consuming and difficult to apply in high-throughput mode, machine learning (ML), a subset of AI, has been widely used to address the question of HS prediction, particularly in the last few years. Both the big boom in data availability as well as more powerful and cheaper software/hardware allowed AI to enhance and accelerate scientific discovery by creating useful knowledge from fragmented information. AI algorithms are very different from traditional analytics as they can analyze much larger datasets and adapt when exposed to new data. Such algorithms have the potential to make accurate predictions given a dataset without needing to be explicitly programmed as they can learn and self-correct to improve their accuracy based on some feedback loop.[195] In fact, their performance improves by learning from previous computations producing reliable and reproducible decisions.

The foundation of ML algorithms is diverse as, for example, some are based on probabilistic models (they model the uncertainty based on probability theory and, in particular, Bayes´ theorem) and others on connectionist approaches (networks of various numerical processors, interconnected and running in parallel such as *artificial neural networks [ANNs]*). They show different behaviors when applied to different scientific problems and as such it is fundamental to test a variety of regression or/and classification algorithm when examining a new biological problem. As the no-free-

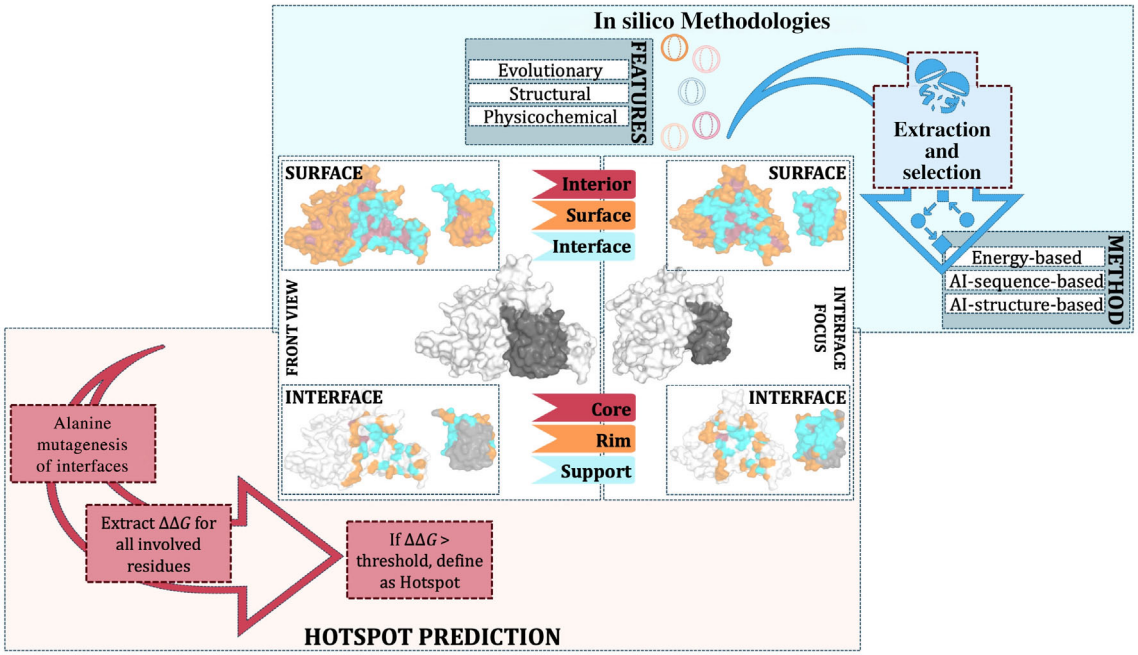

FIGURE 1 HS detection methods workflow: from experimental to in silico methodologies

lunch theorem from Wolpert states: "*The best classifier may not be the same for all datasets*."[196] The "ALGORITHM" column in Table 1 shows that the most used ML algorithms in this field are: *support vector machine* (*SVM*) and *random forest* (*RF*) classifiers. The performance of these methods is usually reported by several threshold-dependent statistical measures derived from a confusion matrix where TP stands for true positive (predicted HS that are actual HS), FP for false positive (predicted HS that are not actual HS), FN for false negative (nonpredicted HS that are actual HS), and TN for true negatives (correctly predicted null-spots). The frequently used metrics are accuracy (ACC), true positive rate (TPR, also called recall or sensitivity), true negative rate (TNR, specificity, or selectivity), positive predictive value (PPV or precision), Matthew's correlation coefficient (MCC), and F-score (F1). Most authors do not publish all available metrics, hampering a proper performance comparison between existing algorithms. Still, Table 1 lists the most relevant metrics for the available algorithms to facilitate the understanding of current state-of-the-art (SOTA) tools.

In principle, a prediction tool should follow several specific procedures to ensure maximum confidence and performance through a multistep implementation:

1. Construction of a valid benchmark dataset, with a well-though split into training (data used to construct the model) and testing (data used to measure the final model performance) sets;
2. Formulation of a set of key features that show some correlation with the quantity to be analyzed/predicted;
3. Introduction and/or development of a powerful algorithm (or engine);

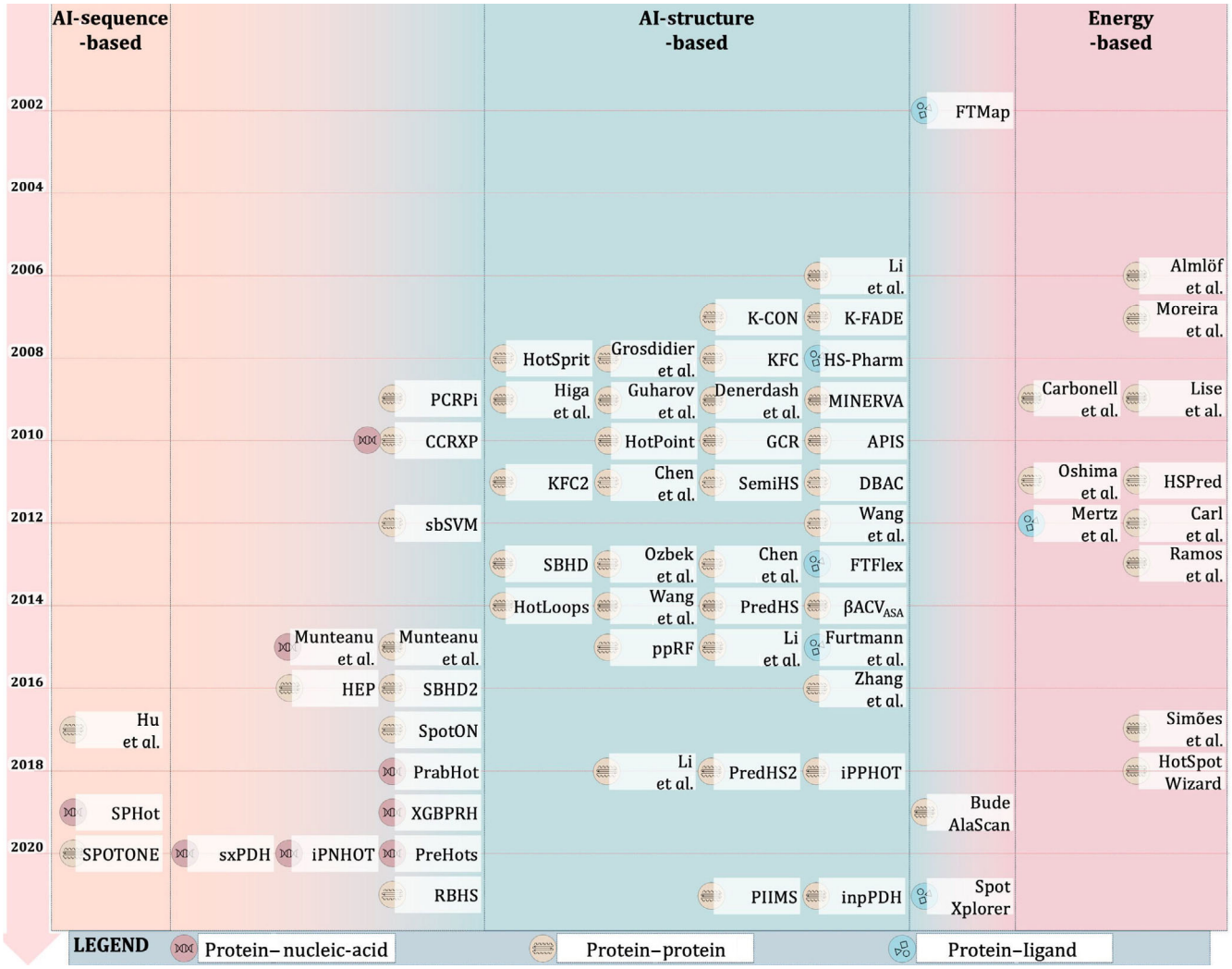

FIGURE 2 Timeline of representative methods developed to HS detection at protein-based complexes (protein–protein, protein–nucleic acid, and protein–ligand). These methods were further split in the three used methodologies: energy-based, AI-sequence-based and AI-structure-based

4. Properly performing *N*-fold or jackknife cross-validation test, subsampling test and independent dataset test to evaluate the performance of the used method; and
5. Ideally, the provision of a user-friendly web-server fully accessible to the public or a simple to use and install version of the software for local use.[109]

The current big data era is resulting in a huge number of rich sources of molecular level information for proteins. These are extremely useful for ML applications which depend on the selection of key features that encode the main characteristics of the biological problem at hands.[197] Usually, researchers will provide a high number of local and global characteristics/features and let the algorithm choose by itself the ones that provide the higher discriminatory power. This learning depends not only on feature correlation but also on their encoding, renormalization and mix between different formats. The increased number of features that can be considered brings however challenges in defining their relative weight.[197]

The used features in the development of different data-driven models typically fall into two broad categories:

1. Sequence-based that use an encoding of sequence-derived features of the residues and their neighbors and explores amino-acid identity, physico-chemical properties of amino-acids, predicted solvent accessibility, position-specific scoring matrices (PSSMs), interface propensities, sequence conservation, and co-evolution; and
2. Structure-based features of the target residues and neighbors such as interface and surface propensities, interface size, geometry, roughness, SASA, atomic interactions, secondary and tertiary structural information, sequence entropy, surface shape, and physico-chemical-based features (amino acid composition and properties, GO-driven frequency-based similarity, and semantic similarity).

As shown, SASA was already reported as a key feature to improve HS detection. It is essential in a wide variety of AI-based models.[75,76,109,131,141,181] However, it has been shown that even conservation and SASA, which were highlighted as key contributors in a binding interface, can alone not unambiguously define a HS.[198] Most developed AI-methods focus on structural features and do not depend on the type of analyzed interface. In contrast to protein–protein interfaces that have been studied for the past 15 years, the development of methods for nucleic acids only took off since 2015. The recent release of new databases of alanine mutations at protein–nucleic acid interfaces will for sure fuel the application of AI-methods to these complexes. Figure 2 illustrates the time evolution of the available algorithms for HS detection.

Applying computational AI methods to high-throughput genomics/proteomics is not a straightforward technical task. In addition to the expected complexity of algorithms, collecting the data, storing them, performing real-time analysis, and distributing the resulting insights are also technical challenges. If enough data are available, a new subset of ML, deep-learning (DL) algorithms could transform HS detection approaches not only as a high-performance prediction tool, but also as a ground-breaking technology. DL is a collection of techniques and methods that are used to build composable differentiable architectures. The more relevant one's for the field are probably *multilayered perceptron* (MLP), *convolutional deep neural networks* (CNNs), *graph convolutional networks* (GCN), and its common variants.[199,200]

DL success in structural biology was recently demonstrated by the development of a neural network-based model, AlphaFold2 (AF2), to accurately predict the 3D structure of a human proteome, among others. Arguably one of the biggest achievements in the structural biology and AI fields, AF2 has demonstrated an exceptional performance in the 14th Critical assessment of protein structure prediction (CASP14) with a median backbone and all-atom accuracy of 0.96 and 1.5 Å root-mean-square-deviation (RMSD), respectively[201] using transformers in an innovative manner since, previously this algorithm was used for image analysis and natural language processing.[202] In fact, AF2 was able to understand complex interrelationships between sequence and structure, and use that information to predict multiple structural features, and ultimately to predict reliable 3D models.[203] Upon the open-source release of AF2, the sizzling scientific community has been publishing encouraging results regarding protein interaction predictions. AF2 was used to assess protein–peptide complex structures and achieved great results with around 40% of the complexes modeled with an accuracy under 2.0 Å (Cα-RMSD) (Ko *et al*., unreviewed results, doi: 10.1101/2021.07.27.453972). Likewise, protein–peptide docking was the focus of the work developed by Schueler-Furman laboratory, which demonstrated that a simpler approach only using AF2 was able to mimic SOTA models with the advantage of the algorithm being simple sequence-based (Tsaban *et al*., unreviewed results, doi: 10.1101/2021.08.01.454656). Elofsson et al. also developed a "fold and dock" pipeline to accurately predict protein–protein complexes. They used AF2 to this end and achieved better results than SOTA software. Adding to the ability to predict the complexes, they could also discriminate between

interacting and noninteracting protein dual sets (Bryant *et al*., unreviewed results, doi: 10.1101/2021.09.15.460468). Despite being a standalone tool, AF2 was also combined with ClusPro[204,205] which increased the success rate by 23% for the top 5% predictions and 40% for the top 10%. Inspired by the release of AlphaFold, before its debut as a public available tool, Baker and co-workers developed and published RoseTTA-Fold.[206] At that time, RoseTTA Fold debut with comparable performance to AlphaFold, the possibility to predict protein–protein complexes, and availability as a public server.[206]

DeepMind took one step further and released AlphaFold-Multimer, a model that aims to predict multichains protein complexes contrasting to their original single-chain structure predictor, AF2. The latest AlphaFold-Multimer could correctly predict 67% and 69% of the heteromeric and homomeric interfaces, respectively, and with high-accuracy predictions in 23% and 34% of the complexes (Evans *et al*., unreviewed results, doi: 10.1101/2021.10.04.463034). To add to the AF2 breakthrough, other groups are contributing with parallel releases encompassing extra features for the original model. Lin and coworkers. developed ParaFold, a solution to improve the central processing unit/graphics processing unit (CPU/GPU) use of AF2 that can be of use to deal with the computational requirements since they were able to speed the predictions almost by 14-fold (Zhong *et al*., unreviewed results, doi: arXiv:2111.06340, 2021). Perrakis et al. developed AlphaFill which adds cofactors and ligands to the AF2 predictions to enhance the biological interpretation (Hekkelman *et al*., unreviewed results, doi: 10.1101/2021.11.26.470110). Skolnick et al. postulate that AF2 models have still to be carefully prepared before used in drug discovery to tackle the protonation state, activation state, presence of ions/solvent, among other relevant factors.[207]

A few authors have also merged characteristics of both methodologies, energy- and AI-based, achieving interesting results. For example, Ibarra et al. developed BudeAlaScan, a consensus ML based function that allows the use of multiple HS detection methods, including energy- and AI-based ones. Moreover, by allowing the upload of nuclear magnetic resonance (NMR) ensembles or MD trajectories, their consensus function exploit intrinsically disordered regions (IDRs) and transient or dynamic noncovalent contacts, further amplifying the potential of HS detection and drug development.[163,208]

5 | CONCLUSION

Despite technological advances, the explosion of genomic and proteomic data and the inherent advances of structural bioinformatics, there is still room to improve the overall performance of HS detection methods. One of the main difficulties lies in the interpretation and mining of an ever-growing, scattered, and overwhelming wealth of diverse data from global systemic approaches with an increased granularity of evidence of which large searchable databases already exist. This deluge of information has also provided us access to a panoply of protein-interactions-related data; a source that remains underexplored. Determining the relative importance of different pieces of evidence when combining the available information to suggest potentially successful binding motifs, and in particular HS, all crucial steps for drug discovery, is another challenge.

Given the latest advances in the field of AI application to structural biology (with AF2 as a recent example), this enormous task can now be pursued. In fact, innovative, fast, and accurate AI procedures are being continuously developed to detect HS at all types of protein-based interfaces. These tools are typically assembled in online, user-friendly platforms, that bridge the gap to the wet-lab, appealing to the scientific community as a less costly and time-effective approach. Undoubtedly, these new techniques are the basis for a new disruptive paradigm in the drug development field.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

AUTHOR CONTRIBUTIONS

Nícia Rosário-Ferreira: Data curation (lead); investigation (equal); visualization (lead); writing – original draft (equal); writing – review and editing (equal). **Alexandre M. J. J. Bonvin:** Conceptualization (equal); investigation (equal); validation (equal); writing – original draft (equal); writing – review and editing (equal). **Irina S. Moreira:** Conceptualization (equal); investigation (equal); project administration (lead); supervision (lead); validation (equal); writing – original draft (equal); writing – review and editing (equal).

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Nícia Rosário-Ferreira https://orcid.org/0000-0002-7225-9287
Alexandre M. J. J. Bonvin https://orcid.org/0000-0001-7369-1322
Irina S. Moreira https://orcid.org/0000-0003-2970-5250

RELATED WIRES ARTICLES

Finding the ΔΔG spot: Are predictors of binding affinity changes upon mutations in protein-protein interactions ready for it?

FURTHER READING

Geng C, Xue LC, Roel-Touris J, Bonvin AMJJ. Finding the ΔΔG spot: are predictors of binding affinity changes upon mutations in protein–protein interactions ready for it? Wiley Interdiscip Rev Comput Mol Sci. 2019;9:e1410.

REFERENCES

1. Stumpf MPH, Thorne T, De Silva E, Stewart R, An HJ, Lappe M, et al. Estimating the size of the human interactome. Proc Natl Acad Sci U S A. 2008;105:6959–64.
2. Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, et al. An empirical framework for binary interactome mapping. Nat Methods. 2009;6:83–90.
3. Moreira IS, Fernandes PA, Ramos MJ. Hot spots: a review of the protein-protein interface determinant amino-acid residues. Proteins Struct Funct Genet. 2007;68:803–12.
4. Jones S, Thornton JM. Principles of protein-protein interactions. Proc Natl Acad Sci U S A. 1996;93:13–20.
5. Ozdemir ES, Gursoy A, Keskin O. Analysis of single amino acid variations in singlet hot spots of protein-protein interfaces. Bioinformatics. 2018;34:i795–801.
6. Alexov E, Sternberg M. Understanding molecular effects of naturally occurring genetic differences. J Mol Biol. 2013;425:3911–3.
7. Yates CM, Sternberg MJE. Proteins and domains vary in their tolerance of non-synonymous single nucleotide polymorphisms (nsSNPs). J Mol Biol. 2013;425:1274–86.
8. Goncearenco A, Li M, Simonetti FL, Shoemaker BA, Panchenko AR. Exploring protein-protein interactions as drug targets for anti-cancer therapy with in silico workflows. Methods Mol Biol. 2017;1647:221–36.
9. Ibarra AA, Bartlett GJ, Hegedüs Z, Dutt S, Hobor F, Horner KA, et al. Predicting and experimentally validating hot-spot residues at protein-protein interfaces. ACS Chem Biol. 2019;14:2252–63.
10. Siebenmorgen T, Zacharias M. Computational prediction of protein–protein binding affinities. Wiley Interdiscip Rev Comput Mol Sci. 2020;10:1–18.
11. Lawson ADG, MacCoss M, Baeten DL, Macpherson A, Shi J, Henry AJ. Modulating target protein biology through the re-mapping of conformational distributions using small molecules. Front Chem. 2021;9:668186.
12. Kunig VBK, Potowski M, Klika Škopić M, Brunschweiger A. Scanning protein surfaces with DNA-encoded libraries. ChemMedChem. 2021;16:1048–62.
13. Sheng C, Dong G, Miao Z, Zhang W, Wang W. State-of-the-art strategies for targeting protein-protein interactions by small-molecule inhibitors. Chem Soc Rev. 2015;44:8238–59.
14. Lu H, Zhou Q, He J, Jiang Z, Peng C, Tong R, et al. Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. Signal Transduct Target Ther. 2020;5:213.
15. Li Z, Ivanov AA, Su R, Gonzalez-Pecchi V, Qi Q, Liu S, et al. The OncoPPi network of cancer-focused protein-protein interactions to inform biological insights and therapeutic strategies. Nat Commun. 2017;8:1–14.
16. Vajda S, Whitty A, Kozakov D. Fragments and hot spots in drug discovery. Oncotarget. 2015;6:18740–1.
17. Clackson T, Wells JA. A hot spot of binding energy in a hormone-receptor interface. Science (80). 1995;267:383–6.
18. Hardy JA, Wells JA. Searching for new allosteric sites in enzymes. Curr Opin Struct Biol. 2004;14:706–15.
19. Arkin MMR, Wells JA. Small-molecule inhibitors of protein-protein interactions: Progressing towards the dream. Nat Rev Drug Discov. 2004;3:301–17.
20. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol. 1998;280:1–9.
21. Chang C, Lin SM, Satange R, Lin SC, Sun SC, Wu HY, et al. Targeting protein-protein interaction interfaces in COVID-19 drug discovery. Comput Struct Biotechnol J. 2021;19:2246–55.
22. Pirolli D, Righino B, de Rosa MC. Targeting SARS-CoV-2 spike protein/ACE2 protein-protein interactions: a computational study. Mol Inform. 2021;40:2060080.
23. Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature. 2020;583:459–68.

24. Villoutreix BO, Calvez V, Marcelin AG, Khatib A. In silico investigation of the new UK (B.1.1.7) and South African (501y.v2) SARS-CoV-2 variants with a focus at the ace2-spike rbd interface. Int J Mol Sci. 2021;22:1–13.
25. Guedes IA, Costa LSC, Dos Santos KB, Karl ALM, Rocha GK, Teixeira IM, et al. Drug design and repurposing with DockThor-VS web server focusing on SARS-CoV-2 therapeutic targets and their non-synonym variants. Sci Rep. 2021;11:5543.
26. Auwul MR, Rahman MR, Gov E, Shahjaman M, Moni MA. Bioinformatics and machine learning approach identifies potential drug targets and pathways in COVID-19. Brief Bioinform. 2021;22:bbab120.
27. Ceylan H. A bioinformatics approach for identifying potential molecular mechanisms and key genes involved in COVID-19 associated cardiac remodeling. Gene Rep. 2021;24:101246.
28. Trigueiro-Louro J, Correia V, Figueiredo-Nunes I, Gíria M, Rebelo-de-Andrade H. Unlocking COVID therapeutic targets: a structure-based rationale against SARS-CoV-2, SARS-CoV and MERS-CoV Spike. Comput Struct Biotechnol J. 2020;18:2117–31.
29. Zahradník J, Schreiber G. Protein engineering in the design of protein–protein interactions: SARS-CoV-2 inhibitors as a test case. Biochemistry. 2021;60:3429–35.
30. Jones S, Thornton JM. Principles of protein-protein interactions. Proc Natl Acad Sci U S A. 1996;93:13–20.
31. Keskin O, Gursoy A, Ma B, Nussinov R. Principles of protein-protein interactions: what are the preferred ways for proteins to interact? Chem Rev. 2008;108:1225–44.
32. Tuncbag N, Gursoy A, Guney E, Nussinov R, Keskin O. Architectures and functional coverage of protein-protein interfaces. J Mol Biol. 2008;381:785–802.
33. Acuner Ozbabacan SE, Engin HB, Gursoy A, Keskin O. Transient proteinprotein interactions. Protein Eng Des Sel. 2011;24:635–48.
34. Rickard MM, Zhang Y, Gruebele M, Pogorelov TV. In-cell protein-protein contacts: transient interactions in the crowd. J Phys Chem Lett. 2019;10:5667–73.
35. Das J, Mohammed J, Yu H. Genome-scale analysis of interaction dynamics reveals organization of biological networks. Bioinformatics. 2012;28:1873–8.
36. Dey S, Pal A, Chakrabarti P, Janin J. The subunit interfaces of weakly associated homodimeric proteins. J Mol Biol. 2010;398:146–60.
37. Perkins JR, Diboun I, Dessailly BH, Lees JG, Orengo C. Transient protein-protein interactions: structural, functional, and network properties. Structure. 2010;18:1233–43.
38. Jayashree S, Murugavel P, Sowdhamini R, Srinivasan N. Interface residues of transient protein-protein complexes have extensive intra-protein interactions apart from inter-protein interactions. Biol Direct. 2019;14:1–14.
39. Yates CM, Sternberg MJE. The effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on protein-protein interactions. J Mol Biol. 2013;425:3949–63.
40. Arkin MR, Randal M, DeLano W, Hyde J, Luong TN, Oslob JD, et al. Binding of small molecules to an adaptive protein-protein interface. Proc Natl Acad Sci U S A. 2003;100:1603–8.
41. Tuncbag N, Gursoy A, Keskin O. Prediction of protein-protein interactions: Unifying evolution and structure at protein interfaces. Phys Biol. 2011;8:035006.
42. Seychell BC, Beck T. Molecular basis for protein–protein interactions. Beilstein J Org Chem. 2021;17:1.
43. Nooren IMA, Thornton JM. Diversity of protein-protein interactions. EMBO J. 2003;22:3486–92.
44. Lins L, Thomas A, Brasseur R. Analysis of accessible surface of residues in proteins. Protein Sci. 2003;12:1406–17.
45. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. Studies of protein-protein interfaces: A statistical analysis of the hydrophobic effect. Protein Sci. 1997;6:53–64.
46. Levy ED. A simple definition of structural regions in proteins and its use in analyzing interface evolution. J Mol Biol. 2010;403:660–70.
47. Kenneth Morrow J, Zhang S. Computational prediction of protein hot spot residues. Curr Drug Metab. 2012;18:1255–65.
48. Wells JA, McClendon CL. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. Nature. 2007;450:1001–9.
49. Conte LL, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol. 1999;285:2177–98.
50. Smith MC, Gestwicki JE. Features of protein-protein interactions that translate into potent inhibitors: topology, surface area and affinity. Expert Rev Mol Med. 2012;14:e16.
51. London N, Raveh B, Schueler-Furman O. Druggable protein-protein interactions: from hot spots to hot segments. Curr Opin Chem Biol. 2013;17:952–9.
52. Chakrabarti P, Janin J. Dissecting protein-protein recognition sites. Proteins Struct Funct Genet. 2002;47:334–43.
53. Guharoy M, Chakrabarti P. Conservation and relative importance of residues across protein-protein interfaces. Proc Natl Acad Sci U S A. 2005;102:15447–52.
54. Navío D, Rosell M, Aguirre J, de la Cruz X, Fernández-Recio J. Structural and computational characterization of disease-related mutations involved in protein-protein interfaces. Int J Mol Sci. 2019;20:1583.
55. David A, Sternberg MJE. The contribution of missense mutations in core and rim residues of protein-protein interfaces to human disease. J Mol Biol. 2015;427:2886–98.
56. Sunami T, Kono H. Local conformational changes in the DNA interfaces of proteins. PLoS One. 2013;8:e56080.
57. Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS. Origins of specificity in protein-DNA recognition. Annu Rev Biochem. 2010;79:233–69.
58. Luscombe NM, Thornton JM. Protein-DNA interactions: Amino acid conservation and the effects of mutations on binding specificity. J Mol Biol. 2002;320:991–1009.
59. Jones S, van Heyningen P, Berman HM, Thornton JM. Protein-DNA interactions: A structural analysis. J Mol Biol. 1999;287:877–96.

60. Janin J, Rodier F, Chakrabarti P, Bahadur RP. Macromolecular recognition in the Protein Data Bank. Acta Crystallogr Sect D Biol Crystallogr. 2006;63:1–8.
61. Gardini S, Furini S, Santucci A, Niccolai N. A structural bioinformatics investigation on protein-DNA complexes delineates their modes of interaction. Mol BioSyst. 2017;13:1010–7.
62. Lejeune D, Delsaux N, Charloteaux B, Thomas A, Brasseur R. Protein-nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure. Proteins Struct Funct Genet. 2005;61:258–71.
63. Schneider B, Cerný J, Svozil D, Cech P, Gelly JC, De Brevern AG. Bioinformatic analysis of the protein/DNA interface. Nucleic Acids Res. 2014;42:3381–94.
64. Pabo CO, Sauer RT. Transcription factors: structural families and principles of DNA recognition. Annu Rev Biochem. 1992;61:1053–95.
65. Kagra D, Prabhakar PS, Sharma KD, Sharma P. Structural patterns and stabilities of hydrogen-bonded pairs involving ribonucleotide bases and arginine, glutamic acid, or glutamine residues of proteins from quantum mechanical calculations. ACS Omega. 2020;5:3612–23.
66. Corsi F, Lavery R, Laine E, Carbone A. Multiple protein-DNA interfaces unravelled by evolutionary information, physico-chemical and geometrical properties. PLoS Comput Biol. 2020;16:e1007624.
67. Petukh M, Kucukkal TG, Alexov E. On human disease-causing amino acid variants: statistical study of sequence and structural patterns. Hum Mutat. 2015;36:524–34.
68. Keskin O, Ma B, Nussinov R. Hot regions in protein-protein interactions: the organization and contribution of structurally conserved hot spot residues. J Mol Biol. 2005;345:1281–94.
69. Li X, Keskin O, Ma B, Nussinov R, Liang J. Protein-protein interactions: Hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: Implications for docking. J Mol Biol. 2004;344:781–95.
70. Chen CZ, Shapiro R. Superadditive and subadditive effects of 'hot spot' mutations within the interfaces of placental ribonuclease inhibitor with angiogenin and ribonuclease A. Biochemistry. 1999;38:9273–85.
71. Moreira IS, Martins JM, Ramos RM, Fernandes PA, Ramos MJ. Understanding the importance of the aromatic amino-acid residues as hot-spots. Biochim Biophys Acta. 2013;1834:404–14.
72. Moza B, Buonpane RA, Zhu P, Herfst CA, Rahman AK, McCormick JK, et al. Long-range cooperative binding effects in a T cell receptor variable domain. Proc Natl Acad Sci U S A. 2006;103:9867–72.
73. Kuttner YY, Engel S. Complementarity of stability patches at the interfaces of protein complexes: implication for the structural organization of energetic hot spots. Proteins Struct Funct Bioinform. 2018;86:229–36.
74. Kuttner YY, Engel S. Protein hot spots: the islands of stability. J Mol Biol. 2012;415:419–28.
75. Moreira IS, Ramos RM, Martins JM, Fernandes PA, Ramos MJ. Are hot-spots occluded from water? J Biomol Struct Dyn. 2014;32:186–97.
76. Moreira I. The role of water occlusion for the definition of a protein binding hot-spot. Curr Top Med Chem. 2015;15:2068–79.
77. Ramos RM, Fernandes LF, Moreira IS. Extending the applicability of the O-ring theory to protein-DNA complexes. Comput Biol Chem. 2013;44:31–9.
78. Ozdemir ES, Halakou F, Nussinov R, Gursoy A, Keskin O. Methods for discovering and targeting druggable protein-protein interfaces and their application to repurposing. Methods Mol Biol. 2019;1903:1–21.
79. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12:56–68.
80. Shirian J, Sharabi O, Shifman JM. Cold spots in protein binding. Trends Biochem Sci. 2016;41:739–45.
81. Gavenonis J, Sheneman BA, Siegert TR, Eshelman MR, Kritzer JA. Comprehensive analysis of loops at protein-protein interfaces for macrocycle design. Nat Chem Biol. 2014;10:716–22.
82. Meireles LMC, Dömling AS, Camacho CJ. ANCHOR: a web server and database for analysis of protein-protein interaction binding pockets for drug discovery. Nucleic Acids Res. 2010;38:W407–11.
83. Hall DR, Kozakov D, Whitty A, Vajda S. Lessons from hot spot analysis for fragment-based drug discovery. Trends Pharmacol Sci. 2015;36:724–36.
84. Zerbe BS, Hall DR, Vajda S, Whitty A, Kozakov D. Relationship between hot spot residues and ligand binding hot spots in protein-protein interfaces. J Chem Inf Model. 2012;52:2236–44.
85. Ding K, Lu Y, Nikolovska-Coleska Z, Qiu S, Ding Y, Gao W, et al. Structure-based design of potent non-peptide MDM2 inhibitors. J Am Chem Soc. 2005;127:10130–1.
86. Brenke R, Kozakov D, Chuang GY, Beglov D, Hall D, Landon MR, et al. Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. Bioinformatics. 2009;25:621–7.
87. Kozakov D, Hall DR, Chuang GY, Cencic R, Brenke R, Grove LE, et al. Structural conservation of druggable hot spots in protein-protein interfaces. Proc Natl Acad Sci U S A. 2011;108:13528–33.
88. Kozakov D, Grove LE, Hall DR, Bohnuud T, Mottarella SE, Luo L, et al. The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. Nat Protoc. 2015;10:733–55.
89. Kozakova D, Hall DR, Jehle S, Luo L, Ochiana SO, Jones EV, et al. Ligand deconstruction: why some fragment binding positions are conserved and others are not. Proc Natl Acad Sci U S A. 2015;112:E2585–94.
90. Defelipe LA, Arcon JP, Modenutti CP, Marti MA, Turjanski AG, Barril X. Solvents to fragments to drugs: MD applications in drug design. Molecules. 2018;23:1–14.
91. Yu W, Lakkaraju SK, Raman EP, Fang L, MacKerell AD Jr. Pharmacophore modeling using site-identification by ligand competitive saturation (SILCS) with multiple probe molecules. J Chem Inf Model. 2015;55:407–20.

92. Metz A, Pfleger C, Kopitz H, Pfeiffer-Marek S, Baringhaus KH, Gohlke H. Hot spots and transient pockets: predicting the determinants of small-molecule binding to a protein-protein interface. J Chem Inf Model. 2012;52:120–33.
93. Bajusz D, Wade WS, Satała G, Bojarski AJ, Ilaš J, Ebner J, et al. Exploring protein hotspots by optimized fragment pharmacophores. Nat Commun. 2021;12:1–10.
94. Dennis S, Kortvelyesi T, Vajda S. Computational mapping identifies the binding sites of organic solvents on proteins. Proc Natl Acad Sci U S A. 2002;99:4290–5.
95. Brenke R, Kozakov D, Chuang GY, Beglov D, Hall D, Landon MR, et al. Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. Bioinformatics. 2009;25:621–7.
96. Barillari C, Marcou G, Rognan D. Hot-spots-guided receptor-based pharmacophores (HS-pharm): A knowledge-based approach to identify ligand-anchoring atoms in protein cavities and prioritize structure-based pharmacophores. J Chem Inf Model. 2008;48:1396–410.
97. Grove LE, Hall DR, Beglov D, Vajda S, Kozakov D. FTFlex: accounting for binding site flexibility to improve fragment-based identification of druggable hot spots. Bioinformatics. 2013;29:1218–9.
98. Furtmann N, Hu Y, Gütschow M, Bajorath J. Identification of interaction hot spots in structures of drug targets on the basis of three-dimensional activity cliff information. Chem Biol Drug Des. 2015;86:1458–65.
99. Pan Y, Zhou S, Guan J. Computationally identifying hot spots in protein-DNA binding interfaces using an ensemble approach. BMC Bioinform. 2020;21:384.
100. Munteanu CR, Pimenta AC, Fernandez-Lozano C, Melo A, Cordeiro MN, Moreira IS. Solvent accessible surface area-based hot-spot detection methods for protein-protein and protein-nucleic acid interfaces. J Chem Inf Model. 2015;55:1077–86.
101. Pan Y, Wang Z, Zhan W, Deng L. Computational identification of binding energy hot spots in protein-RNA complexes using an ensemble approach. Bioinformatics. 2018;34:1473–80.
102. Deng L, Sui Y, Zhang J. Xgbprh: prediction of binding hot spots at protein–rna interfaces utilizing extreme gradient boosting. Genes (Basel). 2019;10:242.
103. Zhu X, Liu L, He J, Fang T, Xiong Y, Mitchell JC. iPNHOT: a knowledge-based approach for identifying protein-nucleic acid interaction hot spots. BMC Bioinform. 2020;211(21):1–24.
104. Li K, Zhang S, Yan D, Bin Y, Xia J. Prediction of hot spots in protein-DNA binding interfaces based on supervised isometric feature mapping and extreme gradient boosting. BMC Bioinform. 2020;21:381.
105. Zhang S, Zhao L, Xia J. SPHot: prediction of hot spots in protein-RNA complexes by protein sequence information and ensemble classifier. IEEE Access. 2019;7:104941–6.
106. Zhang S, Wang L, Zhao L, Li M, Liu M, Li K, et al. An improved DNA-binding hot spot residues prediction method by exploring interfacial neighbor properties. BMC Bioinform. 2021;22:253.
107. Assi SA, Tanaka T, Rabbitts TH, Fernandez-Fuentes N. PCRPi: Presaging Critical Residues in Protein interfaces, a new computational tool to chart hot spots in protein interfaces. Nucleic Acids Res. 2009;38:e86–6.
108. Ahmad S, Keskin O, Mizuguchi K, Sarai A, Nussinov R. CCRXP: Exploring clusters of conserved residues in protein structures. Nucleic Acids Res. 2010;38:W398–401.
109. Moreira IS, Koukos PI, Melo R, Almeida JG, Preto AJ, Schaarschmidt J, et al. SpotOn: high accuracy identification of protein-protein interface hot-spots. Sci Rep. 2017;7:8007.
110. Xu B, Wei X, Deng L, Guan J, Zhou S. A semi-supervised boosting SVM for predicting hot spots at protein-protein Interfaces. BMC Syst Biol. 2012;6:S6.
111. Xia J, Yue Z, di Y, Zhu X, Zheng CH. Predicting hot spots in protein interfaces based on protrusion index, pseudo hydrophobicity and electron-ion interaction pseudopotential features. Oncotarget. 2016;7:18065–75.
112. Melo R, Fieldhouse R, Melo A, Correia JD, Cordeiro MN, Gümüş ZH, et al. A machine learning approach for hot-spot detection at protein-protein interfaces. Int J Mol Sci. 2016;17:1215.
113. Sitani D, Giorgetti A, Alfonso-Prieto M, Carloni P. Robust principal component analysis-based prediction of protein-protein interaction hot spots. Proteins Struct Funct Bioinform. 2021;89:639–47.
114. Chen P, Li J, Wong L, Kuwahara H, Huang JZ, Gao X. Accurate prediction of hot spot residues through physicochemical characteristics of amino acid sequences. Proteins Struct Funct Bioinform. 2013;81:1351–62.
115. Thorn KS, Bogan AA. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinformatics. 2001;17:284–5.
116. Fischer TB, Arunachalam KV, Bailey D, Mangual V, Bakhru S, Russo R, et al. The binding inteference database (BID): a compilation of amino acid hot spots in protein interfaces. Bioinformatics. 2003;19:1453–4.
117. Jankauskaite J, Jiménez-García B, Dapkunas J, Fernández-Recio J, Moal IH. SKEMPI 2.0: An updated benchmark of changes in protein-protein binding energy, kinetics and thermodynamics upon mutation. Bioinformatics. 2019;35:462–9.
118. Hu SS, Chen P, Wang B, Li J. Protein binding hot spots prediction from sequence only by a new ensemble learning method. Amino Acids. 2017;49:1773–85.
119. Preto AJ, Moreira IS. Spotone: Hot spots on protein complexes with extremely randomized trees via sequence-only features. Int J Mol Sci. 2020;21:1–19.
120. Li L, Zhao B, Cui Z, Gan J, Sakharkar MK, Kangueane P. Identification of hot spot residues at protein-protein interface. Bioinformation. 2006;1:121–6.

121. Darnell SJ, Page D, Mitchell JC. An automated decision-tree approach to predicting protein interaction hot spots. Proteins Struct Funct Genet. 2007;68:813–23.
122. Darnell SJ, Page D, Mitchell JC. An automated decision-tree approach to predicting protein interaction hot spots. Proteins Struct Funct Genet. 2007;68:813–23.
123. Darnell SJ, LeGault L, Mitchell JC. KFC Server: interactive forecasting of protein interaction hot spots. Nucleic Acids Res. 2008;36: W265–9.
124. Grosdidier S, Fernández-Recio J. Identification of hot-spot residues in protein-protein interactions by computational docking. BMC Bioinform. 2008;9:447.
125. Guney E, Tuncbag N, Keskin O, Gursoy A. HotSprint: database of computational hot spots in protein interfaces. Nucleic Acids Res. 2007;36:D662–6.
126. Cho KI, Kim D, Lee D. A feature-based approach to modeling protein-protein interaction hot spots. Nucleic Acids Res. 2009;37: 2672–87.
127. Demerdash ONA, Daily MD, Mitchell JC. Structure-based predictive models for allosteric hot spots. PLoS Comput Biol. 2009;5: e1000531.
128. Guharoy M, Chakrabarti P. Empirical estimation of the energetic contribution of individual interface residues in structures of protein-protein complexes. J Comput Aided Mol Des. 2009;23:645–54.
129. Darnell SJ, LeGault L, Mitchell JC. KFC Server: interactive forecasting of protein interaction hot spots. Nucleic Acids Res. 2008;36: 265–9.
130. Higa RH, Tozzi CL. Prediction of binding hot spot residues by using structural and evolutionary parameters. Genet Mol Biol. 2009;32: 626–33.
131. Xia JF, Zhao XM, Song J, Huang DS. APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. BMC Bioinform. 2010;11:1–14.
132. Li Z, Li J. Geometrically centered region: a 'wet' model of protein binding hot spots not excluding water molecules. Proteins Struct Funct Bioinform. 2010;78:3304–16.
133. Tuncbag N, Keskin O, Gursoy A. HotPoint: hot spot prediction server for protein interfaces. Nucleic Acids Res. 2010;38:W402–6.
134. Tuncbag N, Gursoy A, Keskin O. Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. Bioinformatics. 2009;25:1513–20.
135. Li Z, Wong L, Li J. DBAC: a simple prediction method for protein binding hot spots based on burial levels and deeply buried atomic contacts. BMC Syst Biol. 2011;5:S5.
136. Chen R, Zhang Z, Wu D, Zhang P, Zhang X, Wang Y, et al. Prediction of protein interaction hot spots using rough set-based multiple criteria linear programming. J Theor Biol. 2011;269:174–80.
137. Guan J-H, Dong Q-W, Zhou S-G, Deng L. SemiHS: an iterative semi-supervised approach for predicting proteinprotein interaction hot spots. Protein Pept Lett. 2011;18:896–905.
138. Zhu X, Mitchell JC. KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. Proteins Struct Funct Bioinform. 2011;79:2671–83.
139. Wang L, Liu ZP, Zhang XS, Chen L. Prediction of hot spots in protein interfaces using a random forest model with hybrid features. Protein Eng Des Sel. 2012;25:119–26.
140. Ozbek P, Soner S, Haliloglu T. Hot spots in a network of functional sites. PLoS One. 2013;8:e74320.
141. Martins JM, Ramos RM, Pimenta AC, Moreira IS. Solvent-accessible surface area: how well can be applied to hot-spot detection? Proteins Struct Funct Bioinform. 2014;82:479–90.
142. Wang L, Zhang W, Gao Q, Xiong C. Prediction of hot spots in protein interfaces using extreme learning machines with the information of spatial neighbour residues. IET Syst Biol. 2014;8:184–90.
143. Deng L, Guan J, Wei X, Yi Y, Zhang QC, Zhou S. Boosting prediction performance of protein-protein interaction hot spots by using structural neighborhood properties. J Comput Biol. 2013;20:878–91.
144. Deng L, Zhang QC, Chen Z, Meng Y, Guan J, Zhou S. PredHS: A web server for predicting protein-protein interaction hot spots by using structural neighborhood properties. Nucleic Acids Res. 2014;42:W290–5.
145. Liu Q, Hoi SC, Kwoh CK, Wong L, Li J. Integrating water exclusion theory into β contacts to predict binding free energy changes and binding hot spots. BMC Bioinform. 2014;15:57.
146. Li Z, He Y, Wong L, Li J. Burial level change defines a high energetic relevance for protein binding interfaces. IEEE/ACM Trans Comput Biol Bioinform. 2015;12:410–21.
147. Liu Q, Ren J, Song J, Li J. Co-occurring atomic contacts for the characterization of protein binding hot spots. PLoS One. 2015;10: e0144486.
148. Zhang H, Jiang T, Shan G. Identification of hot spots in protein structures using Gaussian network model and Gaussian naive bayes. Biomed Res Int. 2016;2016:1–9.
149. Qiao Y, Xiong Y, Gao H, Zhu X, Chen P. Protein-protein interface hot spots prediction based on a hybrid feature selection strategy. BMC Bioinform. 2018;19:14.
150. Lin X, Zhang X. Prediction of hot regions in PPIs based on improved local community structure detecting. IEEE/ACM Trans Comput Biol Bioinform. 2018;15:1470–9.

151. Wang H, Liu C, Deng L. Enhanced prediction of hot spots at protein-protein interfaces using extreme gradient boosting. Sci Rep. 2018; 8:14285.
152. Wu F-X, Yang JF, Mei LC, Wang F, Hao GF, Yang GF. PIIMS server: a web server for mutation hotspot scanning at the protein–protein interface. J Chem Inf Model. 2021;61:14–20.
153. Almlöf M, Åqvist J, Smalås AO, Brandsdal BO. Probing the effect of point mutations at protein-protein interfaces with free energy calculations. Biophys J. 2006;90:433–42.
154. Lise S, Archambeau C, Pontil M, Jones DT. Prediction of hot spot residues at protein-protein interfaces by combining machine learning and energy-based methods. BMC Bioinform. 2009;10:365.
155. Carbonell P, Nussinov R, del Sol A. Energetic determinants of protein binding specificity: insights into protein interaction networks. Proteomics. 2009;9:1744–53.
156. Lise S, Buchan D, Pontil M, Jones DT. Predictions of hot spot residues at protein-protein interfaces using support vector machines. PLoS One. 2011;6:e16774.
157. Moreira IS, Fernandes PA, Ramos MJ. Computational alanine scanning mutagenesis: an improved methodological approach. J Comput Chem. 2007;28:644–54.
158. Simões ICM, Costa IP, Coimbra JT, Ramos MJ, Fernandes PA. New parameters for higher accuracy in the computation of binding free energy differences upon alanine scanning mutagenesis on protein-protein interfaces. J Chem Inf Model. 2017;57:60–72.
159. Sumbalova L, Stourac J, Martinek T, Bednar D, Damborsky J. HotSpot Wizard 3.0: web server for automated design of mutations and smart libraries based on sequence input information. Nucleic Acids Res. 2018;46:W356–62.
160. Oshima H, Yasuda S, Yoshidome T, Ikeguchi M, Kinoshita M. Crucial importance of the water-entropy effect in predicting hot spots in protein-protein complexes. Phys Chem Chem Phys. 2011;13:16236–46.
161. Carl N, Hodošček M, Vehar B, Konc J, Brooks BR, Janežič D. Correlating protein hot spot surface analysis using ProBiS with simulated free energies of protein-protein interfacial residues. J Chem Inf Model. 2012;52:2541–9.
162. Wood CW, Heal JW, Thomson AR, Bartlett GJ, Ibarra AÁ, Brady RL, et al. ISAMBARD: an open-source computational environment for biomolecular analysis, modelling and design. Bioinformatics. 2017;33:3043–50.
163. Ibarra AA, Bartlett GJ, Hegedüs Z, Dutt S, Hobor F, Horner KA, et al. Predicting and experimentally validating hot-spot residues at protein–protein interfaces. ACS Chem Biol. 2019;14:2252–63.
164. Li Y, Liu Z, Han L, Li C, Wang R. Mining the characteristic interaction patterns on protein-protein binding interfaces. J Chem Inf Model. 2013;53:2437–47.
165. Rosell M, Fernández-Recio J. Hot-spot analysis for drug discovery targeting protein-protein interactions. Expert Opin Drug Discov. 2018;13:327–38.
166. Metz A, Pfleger C, Kopitz H, Pfeiffer-Marek S, Baringhaus KH, Gohlke H. Hot spots and transient pockets: predicting the determinants of small-molecule binding to a protein-protein interface. J Chem Inf Model. 2012;52:120–33.
167. Perkins JR, Diboun I, Dessailly BH, Lees JG, Orengo C. Transient protein-protein interactions: structural, functional, and network properties. Structure. 2010;18:1233–43.
168. Renault L, Guibert B, Cherfils J. Structural snapshots of the mechanism and inhibition of a guanine nucleotide exchange factor. Nature. 2003;426:525–30.
169. Pommier Y, Cherfils J. Interfacial inhibition of macromolecular interactions: nature's paradigm for drug discovery. Trends Pharmacol Sci. 2005;26:138–45.
170. Macalino SJY, Basith S, NAB C, Chang H, Kang S, Choi S. Evolution of in silico strategies for protein-protein interaction drug discovery. Molecules. 2018;23:1963.
171. Rosell M, Fernández-Recio J. Docking-based identification of small-molecule binding sites at protein-protein interfaces. Comput Struct Biotechnol J. 2020;18:3750–61.
172. Kumar MDS, Gromiha MM. PINT: protein-protein interactions thermodynamic database. Nucleic Acids Res. 2006;34:D195.
173. Cho KI, Kim D, Lee D. A feature-based approach to modeling protein-protein interaction hot spots. Nucleic Acids Res. 2009;37:2672–87.
174. Jemimah S, Yugandhar K, Michael Gromiha M. PROXiMATE: a database of mutant protein-protein complex thermodynamics and kinetics. Bioinformatics. 2017;33:2787–8.
175. Liu L, Xiong Y, Gao H, Wei DQ, Mitchell JC, Zhu X. DbAMEPNI: a database of alanine mutagenic effects for protein-nucleic acid interactions. Database. 2018;2018:1–7.
176. Kumar MDS, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, et al. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res. 2006;34:D204–6.
177. Liu J, Liu S, Liu C, Zhang Y, Pan Y, Wang Z, et al. Nabe: an energetic database of amino acid mutations in protein-nucleic acid binding interfaces. Database. 2021;2021:1–8.
178. Hashemi ZS, Zarei M, Fath MK, Ganji M, Farahani MS, Afsharnouri F, et al. In silico approaches for the design and optimization of interfering peptides against protein–protein interactions. Front Mol Biosci. 2021;8:1–25.
179. Jorgensen WL, Thomas LL. Perspective on free-energy perturbation calculations for chemical equilibria. J Chem Theory Comput. 2008; 4:869–76.
180. Martins SA, Perez MA, Moreira IS, Sousa SF, Ramos MJ, Fernandes PA. Computational alanine scanning mutagenesis: MM-PBSA vs TI. J Chem Theory Comput. 2013;9:1311–9.

181. Moreira IS, Fernandes PA, Ramos MJ. Hot spot occlusion from bulk water: a comprehensive study of the complex between the lysozyme HEL and the antibody FVD1.3. J Phys Chem B. 2007;111:2697–706.
182. Moreira IS, Fernandes PA, Ramos MJ. Unraveling the importance of protein-protein interaction: application of a computational alanine-scanning mutagenesis to the study of the IgG1 streptococcal protein G (C2 fragment) complex. J Phys Chem B. 2006;110:10962–9.
183. Moreira IS, Fernandes PA, Ramos MJ. Hot spot computational identification: application to the complex formed between the hen egg white lysozyme (HEL) and the antibody HyHEL-10. Int J Quantum Chem. 2007;107:299–310.
184. Moreira IS, Fernandes PA, Ramos MJ. Computational alanine scanning mutagenesis: an improved methodological approach. J Comput Chem. 2007;28:644–54.
185. Ramos RM, Moreira IS. Computational alanine scanning mutagenesis-an improved methodological approach for protein-DNA complexes. J Chem Theory Comput. 2013;9:4243–56.
186. Huo S, Massova I, Kollman PA. Computational alanine scanning of the 1:1 human growth hormone-receptor complex. J Comput Chem. 2002;23:15–27.
187. Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L, et al. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. Acc Chem Res. 2000;33:889–97.
188. Huang D, Qi Y, Song J, Zhang JZH. Calculation of hot spots for protein–protein interaction in p53/PMI-MDM2/MDMX complexes. J Comput Chem. 2019;40:1045–56.
189. Moreira IS, Fernandes PA, Ramos MJ. Unravelling hot spots: a comprehensive computational mutagenesis study. Theor Chem Acc. 2007;117:99–113.
190. Li C, Jia Z, Chakravorty A, Pahari S, Peng Y, Basu S, et al. DelPhi Suite: new developments and review of functionalities. J Comput Chem. 2019;40:2502–8.
191. Mongan J, Simmerling C, McCammon J, Case DA, Onufriev A. Generalized born model with a simple, robust molecular volume correction. J Chem Theory Comput. 2007;3:156–69.
192. Petukh M, Li M, Alexov E. Predicting binding free energy change caused by point mutations with knowledge-modified MM/PBSA method. PLoS Comput Biol. 2015;11:e1004276.
193. Peng Y, Sun L, Jia Z, Li L, Alexov E. Predicting protein-DNA binding free energy change upon missense mutations using modified MM/PBSA approach: SAMPDI webserver. Bioinformatics. 2018;34:779–86.
194. Qiu L, Yan Y, Sun Z, Song J, Zhang JZH. Interaction entropy for computational alanine scanning in protein–protein binding. Wiley Interdiscip Rev Comput Mol Sci. 2018;8:e1342.
195. Chakraborty A, Mitra S, De D, Pal AJ, Ahmadian A, Ferrarra M. Determining protein-protein interaction using support vector machine: a review. IEEE Access. 2021;9:12473–90.
196. Meester R. Simulation of biological evolution and the NFL theorems. Biol Philos. 2009;24:461–72.
197. Bouvier B. Protein-protein interface topology as a predictor of secondary structure and molecular function using convolutional deep learning. J Chem Inf Model. 2021;61:3292–303.
198. DeLano WL. Unraveling hot spots in binding interfaces: Progress and challenges. Curr Opin Struct Biol. 2002;12:14–20.
199. Lecun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
200. Derevyanko G, Grudinin S, Bengio Y, Lamoureux G. Deep convolutional networks for quality assessment of protein folds. Bioinformatics. 2018;34:4046–53.
201. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–9.
202. Eisenstein M. Artificial intelligence powers protein-folding predictions. Nature. 2021;599:706–8.
203. Pearce R, Zhang Y. Deep learning techniques have significantly impacted protein structure prediction and protein design. Curr Opin Struct Biol. 2021;68:194–207.
204. Vajda S, Yueh C, Beglov D, Bohnuud T, Mottarella SE, Xia B, et al. New additions to the ClusPro server motivated by CAPRI. Proteins Struct Funct Bioinform. 2017;85:435–44.
205. Kozakov D, Hall DR, Xia B, Porter KA, Padhorny D, Yueh C, et al. The ClusPro web server for protein-protein docking. Nat Protoc. 2017;12:255–78.
206. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021;373:871–6.
207. Skolnick J, Gao M, Zhou H, Singh S. AlphaFold 2: why it works and its implications for understanding the relationships of protein sequence, structure, and function. J Chem Inf Model. 2021;61:4827–31.
208. Wood CW, Ibarra AA, Bartlett GJ, Wilson AJ, Woolfson DN, Sessions RB. BAlaS: Fast, interactive and accessible computational alanine-scanning using BudeAlaScan. Bioinformatics. 2020;36:2917–9.

How to cite this article: Rosário-Ferreira N, Bonvin AMJJ, Moreira IS. Using machine-learning-driven approaches to boost hot-spot's knowledge. WIREs Comput Mol Sci. 2022;12:e1602. https://doi.org/10.1002/wcms.1602