Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in
Structural Biology

Enzyme active sites: Identification and prediction of function using computational chemistry

Kelly K. Barnsley[1] and Mary Jo Ondrechen[1,2]

Abstract

Understanding the biochemically active amino acids in proteins is a key factor to improve the knowledge of how enzymes work, to predict the function of newly discovered protein structures of unknown function, and to establish design principles for enzyme engineering. Here, we explore recently reported computational chemistry-based methods for the prediction of active amino acids in protein 3D structures, including biochemically important distal residues, and their implications for functional genomics, for enzyme design, and for enhancing understanding of the function of enzymes.

Addresses

[1] Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA, 02115, USA
[2] Department of Bioengineering, Northeastern University, Boston, MA, 02115, USA

Corresponding authors: Ondrechen, Mary Jo (mjo@neu.edu);
Barnsley, Kelly K (barnsley.k@northeastern.edu)
(Ondrechen M.J.)

Current Opinion in Structural Biology 2022, **74:**102384

This review comes from a themed issue on **Biophysical Methods**

Edited by **David Fushman** and **Dagmar Ringe**

For complete overview of the section, please refer the article collection - Biophysical Methods

Available online 6 May 2022

https://doi.org/10.1016/j.sbi.2022.102384

Introduction

Understanding the biochemically active amino acids in proteins is a key factor to improve knowledge of how enzymes work, to predict the function of newly discovered protein structures of unknown function, and to establish design principles for enzyme engineering. A key question is what gives rise to the biochemical activity of active site amino acids and co-factors? While the side chains of the ionizable residues are merely week acids and bases for the free amino acids in solution, how do they become strongly reactive in the active site pocket? Local effects on biochemically active residues have been discussed in a recent review by Mazmanian, Sargsyan, and Lim [1] and by Coulther, Ko, and Ondrechen (CKO) [2]. Here, we explore methods based on computational chemistry for the prediction of active amino acids in protein 3D structures and their implications for functional genomics, for enzyme design, and for enhancing the understanding of how enzymes achieve their biochemical activity.

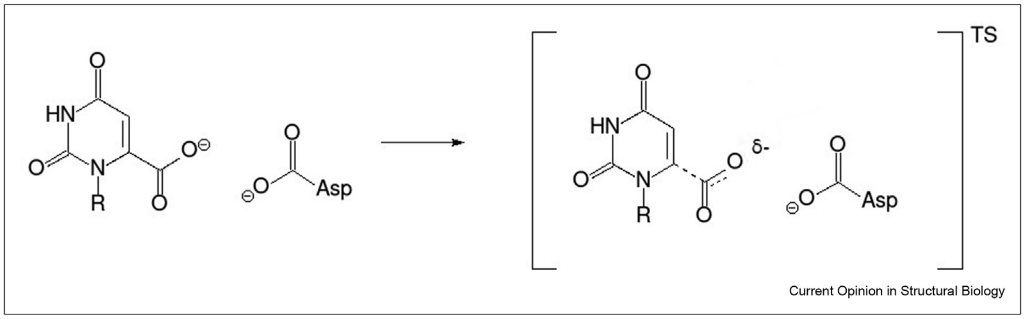
Importance of electrostatic effects

Considerable evidence has been reported that electrostatic effects are very important in enzyme catalysis. Furthermore these effects arise from both adjacent residues and distal residues. Bajorath et al. reported on a density functional theory study comparing free dihydrofolate (DHF) and DHF bound to the active site of a model of the enzyme dihydrofolate reductase. They showed that when the electrical potential of the enzyme is applied to DHF, the electrons in the π orbitals of the reactive C=N double bond of DHF split apart, forming two lobes that more closely resemble atomic 2p orbitals on the C and N atoms. These two lobes are then available to form σ bonds with the 1s orbitals of hydrogen atoms, thus facilitating the reduction reaction. Both nearby and more distant amino acids were reported to contribute to the polarizing potential [3].

Local electric fields inside enzyme active sites can be measured using vibrational Stark spectroscopy and have been shown to be intense [4]. Boxer et al. have reported examples where these electric fields are on the order of 100 MV/cm.

Very recently Chen et al. [5] reported on a theoretical study of the effects of electrostatic interactions on reaction energy barriers in enzymes. The authors argue that energy barriers are lowered by increasing negative charge (or decreasing positive charge) in the electron-donating centers and by increasing positive charge (or decreasing negative charge) in the electron-accepting centers of the reaction. One reported example is the decarboxylation of orotidine 5′-monophosphate catalyzed by orotidine 5′-monophosphate decarboxylase [6,7]. In the enzymatic reaction mechanism, repulsion between the negatively charged carboxylate groups of the substrate and the side chain of Asp70 in the enzyme causes electrostatic stress (Figure 1). The energy of the substrate is raised relative to the transition state and thus the energy barrier for the reaction is lowered.

Figure 1



The initial state and transition state of orotidine 5′-monophosphate, as it undergoes decarboxylation by orotidine 5′-monophosphate decarboxylase. Note the increased repulsion between the carboxylate groups of orotidine 5′-monophosphate and Asp70 in the initial state relative to the transition state, as described by Chen et al. [5].

We have developed methods for the prediction of biochemically active amino acids from protein 3D structures using computed chemical and electrostatic properties [8–12]. One chemical property that facilitates catalysis in enzymes is an expanded buffer range of the ionizable residues involved in catalysis. The buffer range for each ionizable residue is calculated from the theoretical titration curve that is in turn calculated from the electrostatic potential [13–16]. Indeed this computed property is such a universal property of biochemically active residues that we have used it successfully to identify the reactive amino acids in protein structures [8–12]. This expanded buffer range is a simple polyprotic acid effect arising from interactions between residues that can transfer protons. The expanded buffer range insures that both protonation states exist over a wide pH range, so that the active amino acids can return to their original protonation state for the next turnover cycle [2]. The buffer range is the reactive part of the titration curve, as it is the only part of the pH range where both protonation states exist in significant population. Given the dynamic nature of enzymes, an amino acid that follows Henderson–Hasselbalch titration behavior could spend too much time in its asymptotic, unreactive region.

A recent article by Franco and Pessôa Filho [17] describes the importance of anomalous titration behavior of amino acids in proteins and reports on a Hill Equation formalism to describe the anomalous titration curve shapes.

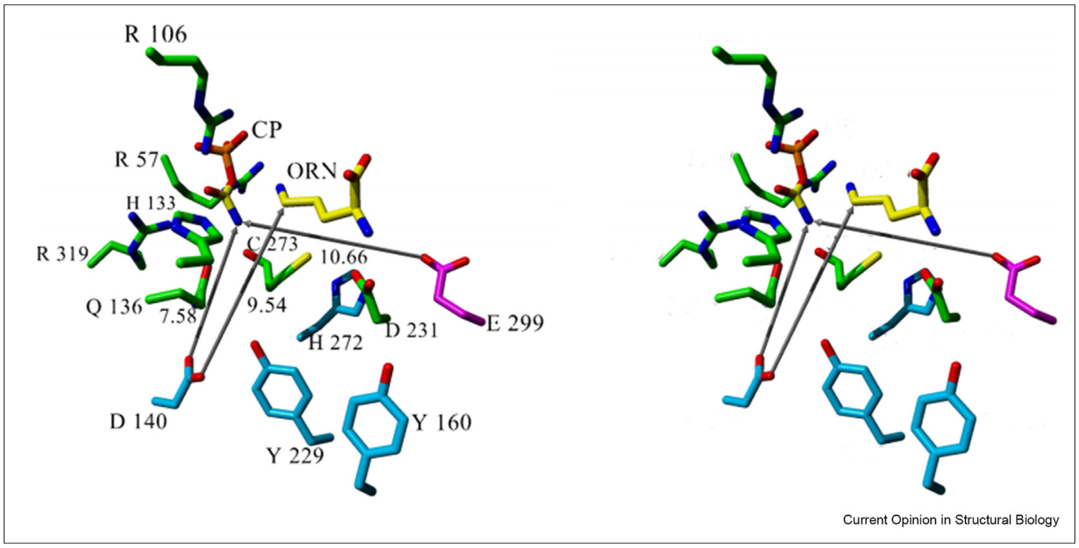
Biochemical roles of distal residues

Reports in the past two years [18,19] have added to the body of evidence that distal residues play key roles in promoting catalytic activity and controlling specificity. Furthermore, these active distal residues are computationally predictable [20].

Escherichia coli prolyl-tRNA synthetase (Ec ProRS) is an aminoacyl-tRNA synthetase whose role is to attach proline to tRNA. Zajac et al. [18] demonstrated that mutations of distal residues not in the main active site have an effect on the catalytic efficiency of Ec ProRs. Two examples are the leucines L281 and L304, whose mutation lowered the proline activation and the efficiency of aminoacylation of the cognate tRNA. The variants T199A, H208A, and V411A were observed to have significantly decreased catalytic efficiency compared to wild type. Mutations in the hinge residues A238 and V391 also have an effect on efficiency, since the hinges are responsible for maintaining the balance of flexibility and stability, which gives the enzyme its catalytic activity.

Ornithine transcarbamoylase (OTC) is an essential enzyme in the urea cycle and arginine biosynthesis pathway; it catalyzes the formation of citrulline (CIT) and inorganic phosphate (P_i) from carbamoyl phosphate (CP) and L-ornithine (ORN) [21–24]. Ngu et al. reported a study of computationally predicted distal residues in *E. coli* OTC, employing site-directed mutagenesis at predicted distal positions, with kinetics and binding assays of wild type and variants, and have established that distal residues play roles in catalytic activity [19]. D140 is a second-shell residue located just behind first-shell residue Q136 with respect to the substrates, about 8 Å from CP and about 10 Å from ORN (Figure 2). The D140N variant has a 28-fold decrease in catalytic efficiency with respect to ORN, compared to wild type. This somewhat lower overall activity suggests that the protonation equilibrium on the side chain of D140, which is lost upon mutation to Asn, couples to the first-shell residues and helps them to affect catalysis. Indeed, results of multi-conformer continuum electrostatics (MCCE) [14,25–27] calculations show that D140 is electrostatically coupled to key first-layer residues R57 and C273. H272 is a second-shell residue, located behind first-shell residues D231 and C273 with respect to ORN. The variant H272L shows a 120-fold decrease in catalytic efficiency with respect to ORN and a 310-fold decrease with respect to CP. E299 is a

Figure 2



Current Opinion in Structural Biology

The active site of *E. coli* ornithine transcarbamylase in cross-eyed stereo. The substrate ORN is shown in yellow. R57, R106, H133, Q136, D231, C273, and R319 are first-shell residues previously reported to be catalytically important [66]. D140, H272, and E299 are distal residues not in direct contact with the substrate that were shown by Ngu et al. [19] to be important for catalysis. Image drawn from structure PDB 1DUV [24] and rendered in YASARA [67].

third-shell residue located behind H272. OTC E299Q exhibits a 51-fold decrease in catalytic efficiency with respect to ORN and a 110-fold decrease with respect to CP. Thus, the protonation equilibrium on the side chain of E299, although not in direct contact with any first-shell residue, contributes to the catalytic reaction located more than 10 Å away (Figure 2).

Yao and Hamelberg [28] have developed a computational approach to the identification of allosteric pathways, wherein networks of interacting residues affect communication across distance within enzymes. Their method is based on difference contact network analysis (dCNA). For the case of imidazole glycerol phosphate synthase (IGPS) [29], in which the catalytic site and the allosteric binding site are 30 Å apart, their approach is able to identify the experimentally-verified residues in the allosteric network. The method is based on conformational ensembles obtained from molecular dynamics simulations starting from both conformational states, that is, the states with and without the allosteric ligand bound.

Javier Garcia-Marin [30] built homology models for two enzymes, the protein tyrosine phosphatase 1B (PTP1B) and lymphocyte T tyrosine phosphatase (TCPTP), involved in Type 2 diabetes, for their apo and inhibitor-bound forms. He produced four 50-ns simulations of both enzymes in their apo and bound states (with the benzbromarone inhibitor BB2 as the ligand in the allosteric site) and compared their RMSD, RMSF, and free energies. This work identifies a phenylalanine, F280, as important to allosteric binding to PTP1B and the likely basis for selective inhibition. It was found that the BB2 ligand shifted position in the binding site of TCPTP but maintained its pose in PTP1B.

Predicting protein function from 3D structure

Structural Genomics work over the past 22 years has produced 15,000+ new three-dimensional structures of proteins [31,32]. Most of these structures are of unknown or uncertain biochemical function. The most commonly used approaches for the prediction of function from structure are informatics-based [33,34]. However computational chemistry-based methods utilizing the local structure at the sites of biochemical activity can be highly effective in the prediction of protein function [35–37].

The ability to identify the biochemically active amino acids in a protein 3D structure can provide information about protein function via local structure matching [37–43]. Bittrich et al. have reported a new method for local structure matching using an inverse indexing approach [44]. Their inverted index compiles all structures in the Protein Data Bank (PDB) that contain specific residue pairs. A local structure motif is described as a set of residue pairs, where residues may be amino acids or nucleic acids. Then all structures in the PDB that contain the specified set of residue pairs may be obtained. Fast computational time is a major advantage of this approach.

The members of the crotonase superfamily catalyze a wide range of reaction types, including important reactions in fatty acid metabolism, with the common feature of a stabilized enolate intermediate [45,46]. In a study of Structural Genomics (SG) proteins in the crotonase superfamily, Mills and et al. [47] used local structural matching [37] of predicted active residues to predict biochemical function. While enzymes in this superfamily often catalyze more than one reaction [48,49], kinetics studies showed that three SG proteins previously annotated as enoyl-CoA hydratases (ECH), but predicted by local structure matching to be hydrolases, have higher catalytic efficiency for hydrolase activity than for ECH activity: Q5SLS5 from *Thermus thermophilus* (PDB 1WZ8), A0A0H2ZRU2 from *Mycobacterium avium* (PDB 3Q1T), and B2HM22 from *Mycobacterium marinum* (PDB 3QK8) were reported to have kinetics similar to that of a previously identified hydrolase, *Cyanobacterium anabaena* β-diketone hydrolase (UniproQ8YNV6, PDBs 2J5S and 2J5G) [50]. Five additional SG proteins were predicted by matching of predicted active residues to be ECHs: Q5KYF9 from *Geobacillus kaustophilus* (PDB 2PPY); Q82Q85 from *Streptomyces avermitilis* (PDB 3H0U); Q5KYB2 from *G. kaustophilus* (PDB 2PBP); Q82QL3 from *S. avermitilis* (PDB 3GKB); and Q6N8W7 from *Rhodopseudomonas palustris* (PDB 3HIN). All five were reported to have ECH activity [47].

The *ab initio* prediction of protein 3D structure from sequence has been hailed as the Method of the Year for 2021 [51]. Indeed the advent of more reliable methods [52,53] using deep learning technologies to predict 3D structure opens the door to the subsequent prediction of function for the vast array of proteins without functional annotations, many newly discovered from gene sequencing. Wehrspan et al. [54] reported identification of iron-sulfur cluster sites and Zn binding sites in the space of over 360,000 protein structures from the AlphaFold database [55]. Those authors report tens of thousands of Fe—S cluster or Zn-binding proteins that are not annotated in Uniprot as such, greatly increasing the set of known metalloproteins.

Feehan, Franklin, and Slusky (FFS) describe a newly developed model, metal activity heuristic of metalloprotein and enzymatic sites (MAHOMES) to distinguish between enzymatic and non-enzymatic metal species in proteins [56]. Their machine learning approach employs multiple input features, including energy terms, pocket geometry, descriptors of adjacent residues, electrostatic metrics, and metal coordination geometry. FFS report that the most important input feature for prediction is an electrostatic metric derived from the theoretical titration curves [9] of the ionizable residues. MAHOMES is reported to predict enzymatic metal species with 90.1% recall and 92.2% precision.

Protein design

The enzymatic production of biofuels from cellulose is one possible path toward sustainable, carbon-recycling fuel production. Summers et al. [57] have used the substitution of both first-shell residues and distal residues to address the problem of cellulase inhibition by its reaction products, a major barrier to the deployment of enzymatic biofuel synthesis. Their designed variants W212A, W213A, Q247A, W249A and F250A of endoglucanase 1 (E1) from *Acidothermus cellulolyticus* [58] are reported to have reduced cellobiose inhibition. W212, W213, and Q247 are first-shell residues, while W249 and F250 are distal residues with side chains located 8—10 Å away from the ligand (Figure 3). Replacement of these residues by alanine induces conformational changes that decrease binding affinity for products and therefore can improve enzyme activity.

The designed retroaldolases are among the most highly evolved of designed enzymes. Coulther et al. [59] analyzed a series of designed retroaldolases in the RA95 family of retroaldolases [60—63]. Specifically they reported on the series RA95.5, RA95.5—5, RA95.5—8, and RA95.5—8F, representing the evolutionary trajectory along which activity is increased by orders of magnitude. The authors demonstrate that, as evolution proceeds and higher activities are achieved, the electrostatic couplings between the catalytically active lysine K83 and surrounding amino acids is increased. Y51 is a key coupled residue and the variant Y51F was observed to have decreased activity along the evolutionary trajectory.

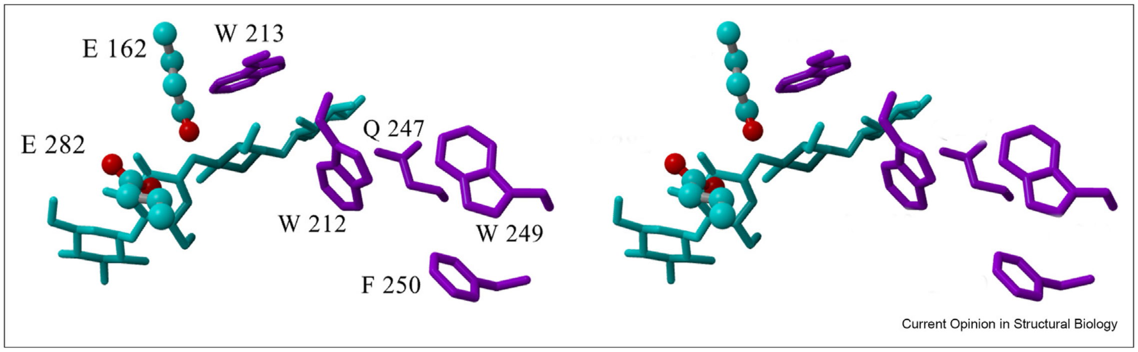
Pirro et al. reported on the design of an allosterically controlled phenol oxidase [64]. The authors recombine a DF (due ferri/two iron) de novo protein with a Zinc-porphyrin (ZnP) binding protein (PS1) to create a new synthetic protein that catalyzes a phenol oxidase reaction like DF but whose catalytic rate can be regulated by ZnP binding with an allosteric site on the PS1 section.

Machine learning methods require a training dataset with reliable labels that is of sufficient size and diversity to make predictions about unlabeled cases. These methods have been used to identify biochemically active residues, to predict protein function, and to predict natural substrates; a 2021 review by Feehan, Montezano, and Slusky provides details about applications of machine learning methods to protein design [65].

Understanding how enzymes work

In the recent article of CKO [2], it is argued that an expanded theoretical buffer range is essentially a universal property of catalytic residues that are ionizable and that specific types of interactions lead to increased buffer ranges. The degree of perturbation of the computed titration curve of a residue depends on the

Figure 3



Current Opinion in Structural Biology

The active site of endoglucanase 1 (E1) from *Acidothermus cellulolyticus*. The ligand, shown in cyan, is a β-(1,4) linked tetramer of D-glucopyranose. The two catalytic glutamates, E162 and E282, are rendered in ball-and-stick form and colored by element. The five residues that, when mutated to alanine, are reported to reduce inhibition by product are shown in magenta. Stereo image drawn from structure PDB 1ECE [58] and rendered in YASARA [67].

electrostatic potential energy ε between the residue and each of its coupling partners and on the difference between their intrinsic pK_as. (The intrinsic pK_a of an amino acid in a protein structure is the pK_a of that residue in the hypothetical state of the protein with all other titratable sites in their charge-neutral state.) Expanded titration curves arise from strong coupling between residues that form like charges when the differences in their intrinsic pK_as are within about 1 pH unit. For pairs of residues that form opposite charges, the intrinsic pK_a of the anion-forming residue must be higher than that of the cation-forming residue to contribute to buffer range expansion. For the case of oppositely-charged residues, the range of intrinsic pK_a differences that lead to expanded buffer ranges are also dependent on the electrostatic potential energy and are given approximately by inequality (1), where ε is in units of −ln(10)RT:

$$\varepsilon - 1 \lesssim pK_{a(intr)anion} - pK_{a(intr)cation} \lesssim \varepsilon + 1 \qquad (1)$$

Thus, catalytic aspartate and glutamate residues tend to be strongly coupled to other aspartate and glutamate residues with similar intrinsic pK_as; they may sometimes be coupled to histidines, where the intrinsic pK_a of the acid is higher than that of the histidine. Catalytic lysines tend to be coupled to tyrosines, where the intrinsic pK_a of the tyrosine is higher than that of the lysine; they may sometimes be coupled to other lysines with intrinsic pK_as within about 1 pH unit. The coupling partners to the catalytic residues may be other first-layer residues or may be distal residues.

Discussion and conclusions

Reliable functional annotation of the thousands of recently discovered protein structures of unknown or uncertain function has yet to be completed. Currently all of the factors that give enzymes their remarkable catalytic power are not yet fully understood. Thus the establishment of the basic principles for engineering novel enzymes remains an unsolved problem. The recent articles discussed here represent significant steps toward the achievement of these difficult but game-changing goals. The computational identification of biochemically active amino acids in proteins enables the prediction of function from structure. The space of 3D structures available for such local-site-based function analysis has been greatly expanded by more reliable *ab initio* structure predictors. The identification of coupled residues, including distal residues, is one key piece necessary for the understanding of how enzymes work and for the development of guiding principles for protein engineering.

Conflict of interest statement

Nothing declared.

Acknowledgments

This work was supported by the National Science Foundation under grant number CHE-1905214. MJO acknowledges Fulbright Research grant 2221103 and thanks Dr. Julianna Oláh, the Hungarian American Fulbright Commission, and the Budapest University of Technology and Economics for hosting her sabbatical leave, when part of this work was done.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest

1. * Mazmanian K, Sargsyan K, Lim C: **How the local environment of functional sites regulates protein function**. *J Am Chem Soc* 2020, **142**:9861–9871.
The authors provide perspective on how local interactions, solvent effects, and conformational flexibility contribute to the biochemical function and regulation of proteins. These effects on the biochemical activity of cysteine residues, including the role of second-shell residues on the metal-binding properties, are presented.

2. ** Coulther TA, Ko J, Ondrechen MJ: **Amino acid interactions that facilitate enzyme catalysis**. *J Chem Phys* 2021, **154**:195101.

The authors present underlying theory for the types of interactions between biochemically active residues and other residues that facilitate catalysis. Specifically, catalytic aspartates and glutamates tend to be coupled to other aspartates and glutamates, while catalytic lysines tend to be coupled to high pK_a, anion-forming residues, tyrosines and cysteines.

3. Bajorath J, Kraut J, Li ZQ, Kitson DH, Hagler AT: **Theoretical studies on the dihydrofolate reductase mechanism: electronic polarization of bound substrates**. *Proc Natl Acad Sci Unit States Am* 1991, **88**:6423–6426.

4. Boxer SG, Fried SD, Schneider SH, Wu Y, Fields Electric, Catalysis Enzyme. In *Catalysis in chemistry and biology, proceedings of the 24th solvay conference on chemistry*. World Scientific Publishing Co.; 2017:274–279.

5. * Chen D, Li Y, Li X, Savidge T, Qian Y, Fan X: **Factors determining the enzyme catalytic power caused by noncovalent interactions: charge alterations in enzyme active sites**. *Arab J Chem* 2022, **15**:103611.

The authors show how energy barriers are lowered in enzyme-catalyzed reactions by increasing negative charge (or decreasing positive charge) in the electron-donating centers and by increasing positive charge (or decreasing negative charge) in the electron-accepting centers of the reaction, providing insight into the local interactions that facilitate catalysis in several different enzymes.

6. Brandão TAS, Richard JP: **Orotidine 5'-monophosphate decarboxylase: the operation of active site chains within and across protein subunits**. *Biochemistry* 2020, **59**:2032–2040.

7. Cristobal JR, Brandão TAS, Reyes AC, Richard JP: **Protein-ribofuranosyl interactions activate orotidine 5'-monophosphate decarboxylase for catalysis**. *Biochemistry* 2021, **60**:3362–3373.

8. Ondrechen MJ, Clifton JG, Ringe D: **THEMATICS: a simple computational predictor of enzyme function from structure**. *Proc Natl Acad Sci U S A* 2001, **98**:12473–12478.

9. Tong W, Wei Y, Murga LF, Ondrechen MJ, Williams RJ: **Partial Order Optimum Likelihood (POOL): maximum likelihood prediction of protein active site residues using 3D structure and sequence properties**. *PLoS Comput Biol* 2009, **5**:e1000266.

10. Somarowthu S, Yang H, Hildebrand DGC, Ondrechen MJ: **High-performance prediction of functional residues in proteins with machine learning and computed input features**. *Biopolymers* 2011, **95**:390–400.

11. Somarowthu S, Ondrechen MJ: **POOL server: machine learning application for functional site prediction in proteins**. *Bioinformatics* 2012, **28**:2078–2079.

12. Ringe D, Wei Y, Boino KR, Ondrechen MJ: **Protein structure to function: insights from computation**. *Cell Mol Life Sci* 2004, **61**: 387–392.

13. Madura JD, Briggs JM, Wade RC, Davis ME, Luty BA, Ilin A, Antosiewicz J, Gilson MK, Bagheri B, Scott LR, *et al.*: **Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics program**. *Comput Phys Commun* 1995, **91**:57–95.

14. Georgescu RE, Alexov EG, Gunner MR: **Combining conformational flexibility and continuum electrostatics for calculating pKa's in proteins**. *Biophys J* 2002, **83**:1731–1748.

15. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA: **Electrostatics of nanosystems: application to microtubules and the ribosome**. *Proc Natl Acad Sci U S A* 2001, **98**:10037–10041.

16. * * Guan X, Leven I, Heidar-Zadeh F, Head-gordon T: **Protein C-GeM: a coarse-grained electron model for fast and accurate protein electrostatics prediction**. *J Chem Inf Model* 2021, **61**: 4357–4369.

The authors developed a new method for charge partitioning that is both accurate and more efficient that extant methods. Called a course-grained electron model (C-GeM), this method does not require either electron density nor electrostatic potentials as input.

17. Franco LFM, Pessoa Filho PA: **Mathematical description of the enzymatic activity of proteins with ionizable groups exhibiting deviations from the henderson-hasselbalch equation**. *Appl Biochem Biotechnol* 2021, https://doi.org/10.1007/s12010-021-03700-y.

18. Zajac J, Anderson H, Adams L, Wangmo D, Suhail S, Almen A, Berns L, Coerber B, Dawson L, Hunger A, *et al.*: **Effects of distal mutations on prolyl-adenylate formation of Escherichia coli prolyl-tRNA synthetase**. *Protein J* 2020, **39**:542–553.

19. Ngu L, Winters JN, Nguyen K, Ramos KE, Delateur NA, Makowski L, Whitford PC, Ondrechen MJ, Beuning PJ: **Probing remote residues important for catalysis in Escherichia coli ornithine transcarbamoylase**. *PLoS One* 2020:15. e0228487.

20. Brodkin HR, DeLateur NA, Somarowthu S, Mills CL, Novak WR, Beuning PJ, Ringe D, Ondrechen MJ: **Prediction of distal residue participation in enzyme catalysis**. *Protein Sci* 2015, **24**:762–778.

21. Kuo LC, Miller AW, Lee S, Kozuma C: **Site-directed mutagenesis of Escherichia coli ornithine transcarbamoylase: role of arginine-57 in substrate binding and catalysis**. *Biochemistry* 1988, **27**:8823–8832.

22. Miller AW, Kuo LC: **Ligand induced isomerizations of Escherichia coli ornithine transcarbamoylase**. *J Biol Chem* 1990, **265**:15023–15027.

23. Couchet M, Breuillard C, Corne C, Rendu J, Morio B, Schlattner U, Moinard C: **Ornithine transcarbamylase - from structure to metabolism: an update**. *Front Physiol* 2021, **12**. 748249–748249.

24. Langley DB, Templeton MD, Fields BA, Mitchell RE, Collyer CA: **Mechanism of inactivation of ornithine transcarbamoylase by Ndelta -(N'-Sulfodiaminophosphinyl)-L-ornithine, a true transition state analogue? Crystal structure and implications for catalytic mechanism**. *J Biol Chem* 2000, **275**: 20012–20019.

25. Gunner MR, Zhu X, Klein MC: **MCCE analysis of the pKas of introduced buried acids and bases in staphylococcal nuclease**. *Proteins* 2011, **79**:3306–3319.

26. Alexov EG, Gunner MR: **Incorporating protein conformational flexibility into the calculation of pH-dependent protein properties**. *Biophys J* 1997, **72**:2075–2093.

27. Song Y, Mao J, Gunner MR: **MCCE2: improving protein pKa calculations with extensive side chain rotamer sampling**. *J Comput Chem* 2009, **30**:2231–2247.

28. * * Yao X-Q, Hamelberg D: **Residue–residue contact changes during functional processes define allosteric communication pathways**. *J Chem Theor Comput* 2022, https://doi.org/10.1021/acs.jctc.1c00669.

The authors developed a new computational approach to the identification of the residues involved in allosteric communication. A key point is that the required conformational ensembles are generated from simulations starting from both conformational states (with and without the allosteric ligand bound).

29. Chaudhuri BN, Lange SC, Myers RS, Davisson VJ, Smith JL: **Toward understanding the mechanism of the complex cyclization reaction catalyzed by imidazole glycerolphosphate synthase: crystal structures of a ternary complex and the free enzyme**. *Biochemistry* 2003, **42**:7003–7012.

30. Javier GM: **Computational insight into the selective allosteric inhibition for PTP1B versus TCPTP: a molecular modelling study**. *J Biomol Struct Dyn* 2021, **39**:5399–5410.

31. Chen YW, Yiu C-PB. *Structural genomics: general applications*. New York: Humana; 2021.

32. Michalska K, Joachimiak A: **Structural genomics and the protein Data Bank**. *J Biol Chem* 2021, **296**:100747.

33. Zhao B, Zhang Z, Jiang M, Hu S, Luo Y, Wang L: **NPF:network propagation for protein function prediction**. *BMC Bioinf* 2020, **21**. 355–355.

34. Seyyedsalehi SF, Soleymani M, Rabiee HR, Mofrad MRK: **PFP-WGAN: protein function prediction by discovering Gene Ontology term correlations with generative adversarial networks**. *PLoS One* 2021:16. e0244430.

35. Meng EC, Polacco BJ, Babbitt PC: **Superfamily active site templates**. *Proteins* 2004, **55**:962–976.

36. Shulman-Peleg A, Nussinov R, Wolfson H: **SiteEngines: recognition and comparison of binding sites and protein-protein interfaces**. *Nucleic Acids Res* 2005, **33**:W337–W341.

37. Wang Z, Yin P, Lee JS, Parasuram R, Somarowthu S, Ondrechen MJ: **Protein function annotation with structurally aligned local sites of activity (SALSAs)**. *BMC Bioinf* 2013, **14**. S13.

38. * Zhao J, Cao Y, Zhang L: **Exploring the computational methods for protein-ligand binding site prediction**. *Comput Struct Biotechnol J* 2020, **18**:417–426.

The authors conducted a review of the published literature on ligand binding site (LBS) prediction methods, including 3D structure-based, template-based, traditional and deep machine learning methods.

39. * Guterres H, Park SJ, Zhang H, Im W: **CHARMM-GUI LBS finder & refiner for ligand binding site prediction and refinement**. *J Chem Inf Model* 2021, **61**:3744–3751.

The authors report on a new CHARMM-GUI application that predicts, characterizes and refines ligand binding sites (LBS). *LBS Finder & Refiner* is special because it enables alignment and comparison of binding sites for function prediction, in addition to uses in structure-based drug discovery.

40. Mills CL, Garg R, Lee JS, Tian L, Suciu A, Cooperman G, Beuning PJ, Ondrechen MJ: **Functional classification of protein structures by local structure matching in graph representation**. *Protein Sci* 2018, **27**:1125–1135.

41. * Guterres H, Park S-J, Cao Y, Im W: **CHARMM-GUI ligand designer for template-based virtual ligand design in a binding site**. *J Chem Inf Model* 2021, **61**:5336–5342.

The authors developed a new program or function for CHARMM-GUI that generates virtual ligands based on a protein's binding site. The authors describe the function and usage of *Ligand Designer*, a web-based application that allows users to generate virtual ligands from a protein's binding site, which can then be used in further experiments including ligand design, free energy calculations and molecular dynamics (MD).

42. Stark A, Sunyaev S, Russell RB: **A model for statistical significance of local similarities in structure**. *J Mol Biol* 2003, **326**: 1307–1316.

43. Kleywegt GJ: **Recognition of spatial motifs in protein structures11Edited by J. Thornton**. *J Mol Biol* 1999, **285**:1887–1897.

44. Bittrich S, Burley SK, Rose AS: **Real-time structural motif searching in proteins using an inverted index strategy**. *PLoS Comput Biol* 2020, **16**:e1008502.

45. Holden HM, Benning MM, Haller T, Gerlt JA: **The crotonase superfamily: divergently related enzymes that catalyze different reactions involving acyl coenzyme a thioesters**. *Acc Chem Res* 2001, **34**:145–157.

46. Hamed RB, Batchelar ET, Clifton IJ, Schofield CJ: **Mechanisms and structures of crotonase superfamily enzymes – how nature controls enolate and oxyanion reactivity**. *Cell Mol Life Sci* 2008, **65**:2507–2527.

47. Mills CL, Yin P, Leifer B, Ferrins L, O'Doherty GA, Beuning PJ, Ondrechen MJ: **Functional characterization of structural genomics proteins in the crotonase superfamily**. *ACS Chem Biol* 2022, https://doi.org/10.1021/acschembio.1c00842.

48. Kasaragod P, Schmitz W, Hiltunen JK, Wierenga RK: **The isomerase and hydratase reaction mechanism of the crotonase active site of the multifunctional enzyme (type-1), as deduced from structures of complexes with 3S-hydroxy-acyl-CoA**. *FEBS J* 2013, **280**:3160–3175.

49. Gerlt JA, Babbitt PC: **Enzyme (re)design: lessons from natural evolution and computation**. *Curr Opin Chem Biol* 2009, **13**:10–18.

50. Bennett JP, Whittingham JL, Brzozowski AM, Leonard PM, Grogan G: **Structural characterization of a beta-diketone hydrolase from the cyanobacterium *Anabaena sp.* PCC 7120 in native and product-bound forms, a coenzyme A-independent member of the crotonase suprafamily**. *Biochemistry* 2007, **46**: 137–144.

51. Method of the Year 2021: **Protein structure prediction**. *Nat Methods* 2022, **19**. 1–1.

52. * * Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, *et al.*: **Highly accurate protein structure prediction with AlphaFold**. *Nature* 2021, **596**:583–589.

The authors describe the development of the AlphaFold structure prediction method. AlphaFold showed substantial improvement over previous methods in the Critical Assessment of Structure Prediction.

53. * * Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee Gyu R, Wang J, Cong Q, Kinch Lisa N, Schaeffer RD, *et al.*: **Accurate prediction of protein structures and interactions using a three-track neural network**. *Science* 2021, **373**: 871–876.

A three-track neural network method for the *ab initio* prediction of protein structure from sequence is reported. Accuracy of predictions approaches those of AlphaFold. This work is significant because it addresses a longstanding problem and because the method is freely available to the scientific community.

54. * Wehrspan ZJ, McDonnell RT, Elcock AH: **Identification of iron-sulfur (Fe-S) cluster and Zinc (Zn) binding sites within proteomes predicted by DeepMind's AlphaFold2 program dramatically expands the metalloproteome**. *J Mol Biol* 2022, **434**:167377.

The authors use the AlphaFold database to predict metal-binding sites and report tens of thousands of Fe–S cluster or Zn-binding proteins that are not annotated in Uniprot as such, substantially increasing the set of known metalloproteins.

55. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, *et al.*: **AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models**. *Nucleic Acids Res* 2022, **50**:D439–D444.

56. * * Feehan R, Franklin MW, Slusky JSG: **Machine learning differentiates enzymatic and non-enzymatic metals in proteins**. *Nat Commun* 2021, **12**:3712.

The authors describe a newly developed model, Metal Activity Heuristic of Metalloprotein and Enzymatic Sites (MAHOMES) to distinguish between enzymatic and non-enzymatic metal species in proteins, wherein the most important predictive input features are properties computed from the electrostatic potential.

57. * Summers SR, Alamdari S, Kraft CJ, Brunecky R, Pfaendtner J, Kaar JL: **substitution of distal and active site residues reduces product inhibition of E1 from Acidothermus Cellulolyticus**. *Protein Eng Des Sel* 2021, **34**. gzab031.

Variants of endoglucanase 1 with mutations in distal positions are reported to exhibit reduced inhibition by reaction products. This work thus addresses a major challenge in the production of biofuels from biomass.

58. Sakon J, Adney WS, Himmel ME, Thomas SR, Karplus PA: **Crystal structure of thermostable family 5 endocellulase E1 from Acidothermus cellulolyticus in complex with cellotetraose**. *Biochemistry* 1996, **35**:10648–10660.

59. Coulther TA, Pott M, Zeymer C, Hilvert D, Ondrechen MJ: **Analysis of electrostatic coupling throughout the laboratory evolution of a designed retroaldolase**. *Protein Sci* 2021, **30**: 1617–1627.

60. Jiang L, Althoff EA, Clemente FR, Doyle L, Röthlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF, *et al.*: **De novo computational design of retro-aldol enzymes**. *Science* 2008, **319**:1387–1391.

61. Althoff EA, Wang L, Jiang L, Giger L, Lassila JK, Wang Z, Smith M, Hari S, Kast P, Herschlag D, *et al.*: **Robust design and optimization of retroaldol enzymes**. *Protein Sci* 2012, **21**: 717–726.

62. Giger L, Caner S, Obexer R, Kast P, Baker D, Ban N, Hilvert D: **Evolution of a designed retro-aldolase leads to complete active site remodeling**. *Nat Chem Biol* 2013, **9**:494–498.

63. Obexer R, Godina A, Garrabou X, Mittl PRE, Baker D, Griffiths AD, Hilvert D: **Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase**. *Nat Chem* 2017, **9**:50–56.

64. Pirro F, Schmidt N, Lincoff J, Widel ZX, Polizzi NF, Liu L, Therien MJ, Grabe M, Chino M, Lombardi A, *et al.*: **Allosteric cooperation in a de novo-designed two-domain protein**. *Proc Natl Acad Sci U S A* 2020, **117**:33246–33253.

65. Feehan R, Montezano D, Slusky JSG: **Machine learning for enzyme engineering, selection and design**. *Protein Eng Des Sel* 2021, **34**. gzab019.

66. Ha Y, McCann MT, Tuchman M, Allewell NM: **Substrate-induced conformational change in a trimeric ornithine-transcarbamoylase**. *Proc Natl Acad Sci Unit States Am* 1997, **94**:9550–9555.

67. Krieger E, Joo K, Lee J, Lee J, Raman S, Thompson J, Tyka M, Baker D, Karplus K: **Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: four approaches that performed well in CASP8**. *Proteins* 2009, **77**:114–122.