Chapter 29

Computational and Experimental Studies of ADP-Ribosylation

Robert G. Hammond, Xuan Tan, Matthew Chan, Anupam Goel, and Margaret A. Johnson

Abstract

The macrodomains are a multifunctional protein family that function as receptors and enzymes acting on poly(ADP-ribose), ADP-ribosylated proteins, and other metabolites of nicotinamide adenine dinucleotide (NAD\(^+\)). Several new functions for macrodomains, such as nucleic acid binding and protein–protein interaction, have recently been identified in this family. Here, we discuss methods for the identification of new macrodomains in viruses and the prediction of their function. This is followed by the expression and purification of these proteins following overexpression in bacterial cells and confirmation of folding and function using biophysical methods.

Key words Macrodomain, Coronavirus, Bioinformatics, Docking, Expression, Purification, NMR spectroscopy, STD-NMR, Circular dichroism

1 Introduction

In 2012, a new and dangerous human virus was identified in travelers to the Middle East and Asia. The virus spread rapidly with a fatality rate of about 40% [1, 2]. The Middle East respiratory syndrome (MERS) virus belongs to the betacoronavirus genus which also hosts the pandemic 2003–2005 severe acute respiratory syndrome (SARS) virus that affected more than 30 countries [3]. These viruses have large, complex positive-sense RNA genomes that encode replicase, structural and accessory proteins. The replicase protein is cleaved by a viral protease to form the 15 or 16 nonstructural proteins [4]. The nonstructural proteins (nsps) are responsible for replication of the positive-sense RNA genome, transcription of subgenomic RNAs, and other RNA processing activities. Activities necessary for interference with host cell innate immune responses have been identified [5, 6].
The nonstructural protein 3 is a multifunctional protein consisting of several functional domains and approximately 2000 amino acid residues. This protein harbors one papain-like cysteine protease, transmembrane regions, RNA-binding proteins, and one or more macrodomains [7–9]. The macrodomains are of significant interest with respect to poly(ADP-ribose) (PAR) interaction. These domains comprise a multifunctional protein family that act on NAD+ metabolites [10, 11]. They act as modulators of posttranslational modifications, including PARylation, mono(ADP-ribosylation), and related modifications. In addition to their known roles in PAR binding, some macrodomains function as enzymes, for example in O-acetyl ADP-ribose deacetylase and mono-ADP-ribose hydrolase reactions [12]. The catalytic domain of poly(ADP-ribose) glycosylase is also a macrodomain [13]. These domains have been termed readers, erasers and interpreters of PARylation and mARylation [10]. Divergent macrodomains have been discovered with new functions, including nucleic acid binding [9, 14] and protein–protein interaction [15], suggesting that this protein family represents a conserved structural scaffold on which much functional variation may occur.

In viruses, the nonstructural protein 3 has several roles. Sequence analysis identified this protein as subject to positive selection during coronavirus evolution and adaptation to new hosts [16]. Other lines of evidence show this protein being involved in inhibition of host innate immune responses [17–19]. The macrodomain of this protein is conserved among coronaviruses, hepatitis virus, alphavirus, and rubella virus [20–22]. This protein is able to dephosphorylate ADP-ribose-1″-phosphate; the importance of this reaction to viral infection is unknown. Some conserved sequence motifs, such as a “GGG” and the “SAGIF” which is involved in the stabilization of the phosphate groups, and “NAAN” motif which employs the second asparagine to form hydrogen bonds with the terminal ribose, are characteristic for this protein family [23]. In some viral macrodomains, the latter asparagine participates in the dephosphorylation reaction [10]. Other functional residues are variable, and more difficult to identify on the basis of sequence [10]. The macrodomain is important to infection through interactions with the host immune system: deletion or mutation of the conserved macrodomain led to a loss of virulence [18, 24–26]. In the severe acute respiratory syndrome coronavirus (SARS-CoV), deletion of either the second macrodomain or the domain of unknown function led to a significant loss of viral RNA synthesis, indicating a critical role in the viral replicase-transcriptase complex [14].

In addition to this conserved macrodomain, many coronaviruses contain one or two additional macrodomains. These proteins lack conserved sequence motifs. Originally identified in the SARS-CoV, a typical structure for this region comprises a conserved
macrodomain followed by two divergent macrodomains and a small domain of unknown function [9]. Based on in vitro studies, these proteins were shown to have guanine quadruplex-binding activity. The C-terminal domain was shown to regulate the specificity of RNA binding of the macrodomains [9]. The region containing divergent macrodomains was termed the “SARS-unique domain” based on its lack of sequence identity to other viruses and known proteins [27].

Here, we present methods for the identification of new macrodomains and associated proteins in viruses and other species, and the prediction of their likely domain boundaries and functions. This is followed by the experimental characterization of these proteins by bacterial overexpression, purification, and biophysical and functional characterization, including binding studies with potential ligands.

2 Materials

2.1 Bioinformatics

1. BLAST: https://blast.ncbi.nlm.nih.gov/Blast.cgi.
2. FFAS, Fold & Function Assignment [28]: http://ffas.sanfordburnham.org/ffas-cgi/ffas.pl (see Note 1).
3. Jpred [29]: http://www.compbio.dundee.ac.uk/jpred/.
4. PSIPRED [30]: http://bioinf.cs.ucl.ac.uk/psipred/.
5. DISOPRED [31]: http://bioinf.cs.ucl.ac.uk/psipred/.
6. SEG [32]: http://mendel.imp.ac.at/METHODS/seg.server.html.
7. DisEMBL™ [33]: http://dis.embl.de/.
8. IUPred [34]: http://iupred.enzim.hu/.
9. PrDOS [35]: http://prdos.hgc.jp/cgi-bin/top.cgi.
10. Protein sequences to be analyzed in raw text or FASTA format.

2.2 Protein Functional Analysis

2.2.1 Functional Analysis Servers and Input Files

1. I-TASSER [36]: http://zhanglab.ccmb.med.umich.edu/I-TASSER/.
2. BindN [37]: http://bioinfo.ggc.org/bindn/.
3. CoFactor [38]: https://www.ebi.ac.uk/thornton-srv/databases/CoFactor/.
4. RaptorX [39, 40]: http://raptorx.uchicago.edu/StructurePrediction/.
5. SIFTER [41, 42]: http://sifter.berkeley.edu/.
6. ProFunc [43]: http://www.ebi.ac.uk/thornton-srv/databases/ProFunc/.
7. **COACH** [44]: http://zhanglab.ccmb.med.umich.edu/COACH/.

8. BioLiP [45]: http://zhanglab.ccmb.med.umich.edu/BioLiP/.

9. Protein sequences to be analyzed in raw text or FASTA format.

### 2.2.2 Computational Docking

1. Files in PDB format representing structures of the protein and ligands to be doMcked.

2. JMol program for molecular graphics visualization, accessible at http://www.jmol.org.

3. ZDOCK program for computational docking, accessible at http://zdock.umassmed.edu/.

4. HADDOCK program for computational docking, Web version accessible at http://haddock.science.uu.nl/services/HADDOCK2.2.

5. Manual for HADDOCK, available at http://www.bonvinlab.org/software/haddock2.2/pdb/.

6. Description of analysis with HADDOCK, accessible at http://www.bonvinlab.org/software/haddock2.2/analysis/.

7. GitHub repository of HADDOCK, available at https://github.com/haddocking/pdb-tools.

8. 3D-DART, http://haddock.science.uu.nl/dna/dna.php.

9. Schrodinger Glide: http://www.schrodinger.com/Glide. (Small-Molecule Drug Discovery Suite 2015-4: Glide, version 6.9, Schrödinger, LLC, New York, NY, 2015.)

### 2.3 Protein Expression and Purification

#### 2.3.1 Rich Media Expression

All solutions are prepared with ultrapurified water using ELGA PURELAB Ultra at a sensitivity of 18.2 Ω/cm at 25 °C. Specified solutions are filtered with Millex® GV syringe filter units: 0.22 μm, PVDF, 33 mm diameter. Solutions are prepared and stored at 4 °C (unless indicated otherwise or to manufacturer’s specifications).

1. LB agar.

2. Petri dishes.

3. Plasmid containing gene of interest (pET-15b-TEV).

4. LB-Miller medium: Dissolve 25 g per 1 L of water and autoclave.

5. Ampicillin (1000×): Weigh 1 g ampicillin per 10 mL of water (100 mg/mL). Mix until dissolved. Filter using syringe filter. Store at 4 °C.

6. Isopropyl β-D-1-thiogalactopyranoside (IPTG): 1 M solution in water. Filter using syringe filter. Aliquot and store at −20 °C.
7. 17 × 100 mm disposable polypropylene 5 mL culture tubes.
8. 500 mL and 2800 mL culture flasks.
9. Disposable inoculating loops—sterile plastic loops.
10. BL21(DE3) competent *E. coli*.
11. Plastic centrifuge bottles with caps—centrifuge tubes.

2.3.2 Minimal Media Expression

Table 1

| Salt                          | Amount   |
|-------------------------------|----------|
| 25 mM FeCl$_2$·4H$_2$O        | 5 g      |
| 1.25 mM CaCl$_2$·2H$_2$O      | 184 mg   |
| 1.03 mM H$_3$BO$_3$           | 64 mg    |
| 75.65 μM CoCl$_2$·6H$_2$O     | 18 mg    |
| 23.46 μM CuCl$_2$·2H$_2$O     | 4 mg     |
| 2.49 mM ZnCl$_2$              | 340 mg   |
| 2.50 mM Na$_2$MoO$_4$·2H$_2$O | 605 mg   |
| 202 μM MnCl$_2$·4H$_2$O       | 40 mg    |
Buffers are filtered with a vacuum filtering unit (0.22 μm) and degassed for 5 min in a sonicator bath before use. Purification is performed with an ÄKTA™ purifier system equipped with a Frac-920 fraction collector, P-960 pump, and 50 mL Superloop™ (GE Healthcare). Columns include a 320 mL HiLoad™ 260/600 Superdex™ 200 PG column and 5 mL Histrap™ FF affinity column. Procedures are controlled with UNICORN™ 5.31 software (GE Healthcare).

1. Running buffer: 20 mM sodium phosphate (pH 7.2), 300 mM NaCl, 3 mM dithiothreitol (DTT), 20 mM imidazole. 33.8 mL of 1 M sodium phosphate dibasic, 6.2 mL of 1 M sodium phosphate monobasic, 35.10 g NaCl, 2.72 g imidazole, 0.92 g DTT in 2 L of water. Filter and store at 4 °C.

2. Elution buffer: 20 mM sodium phosphate (pH 7.2), 300 mM NaCl, 3 mM DTT, 500 mM imidazole. 33.8 mL of 1 M sodium phosphate dibasic, 6.2 mL of 1 M sodium phosphate monobasic, 35.10 g NaCl, 68 g imidazole, 0.92 g DTT in 2 L of water. Filter and store at 4 °C.

3. Lysis buffer: 50 mL of running buffer, 500 μL of Triton™ X-100, 1 Pierce™ EDTA-free protease inhibitor tablet (see Note 3). Stir until dissolved. Filtering and degassing not needed.

4. 1 M stock solution of imidazole, 99%; 68.08 g/L.

5. 1,4-Dithiothreitol (DTT).

6. Misonix Ultrasonic Liquid Processor equipped with Model CL5 converter.

7. Millex® GV syringe filter units: 0.22 μm, PVDF, 33 mm diameter.

8. Snakeskin® dialysis tubing: 3.5K molecular weight cutoff (MWCO), 35 mm diameter.

9. AcTEV™ protease.

10. Gel filtration buffer: 20 mM sodium phosphate (pH 7.2), 300 mM NaCl, 5 mM DTT in 2 L of water. Filter and store at 4 °C.

11. Sartorius™ VIVASPIN™ 20 3000-MWCO polyethersulfone (PES) concentrator.

12. Luer-Lok™ Tip 10 mL syringe—plastic syringe.

2.4 Protein Biophysical Characterization

2.4.1 NMR Spectroscopy

1. Sartorius™ VIVASPIN™ 20 3000-MWCO polyethersulfone (PES) concentrator.

2. NMR buffer: 10 mM sodium phosphate, 150 mM NaCl, pH 7.2.

3. 99.9% Deuterium oxide (D₂O) (Cambridge Isotope Laboratories, Inc.).
4. Deuterated-1,4-Dithiothreitol (DTT-d\(_{10}\), 98%): 5 mM DTT-d\(_{10}\) in D\(_2\)O. Store at −20 °C.
5. Sodium azide: 3% (w/v) solution in water. Store at 4 °C.
6. NMR tube, 7-inch × 5 mm.
7. Extended tip pipets: 9 in.
8. Bruker Avance 700 UltraShieldPlus NMR Spectrometer equipped with a CP TCI H-C/N-D cryoprobe. TopSpin™ 3.2 software. SampleCase sample changer (Bruker).

2.4.2 Circular Dichroism Spectroscopy
1. Jasco J-815 CD Spectrometer.
2. Spectra Manager for Windows 95/NT v.1.54.
3. 0.1 mm quartz cuvette.
4. CD buffer: 10 mM sodium phosphate, 75 mM NaCl, pH 7.2.
5. Purified protein in gel filtration buffer.

2.5 Protein Functional Assignment Experiments
1. Mini-cell electrophoresis system.
2. Novex® 8% TBE gel.
3. Novex® TBE running buffer (1x): Dilute 200 mL of 5x stock into water to make up 1 L.
4. Nuclease-free water.
5. EMSA buffer: 10 mM sodium phosphate, 75 mM NaCl, pH 7.2.
6. Oligonucleotide sequences: 5 μM in EMSA buffer.
7. Novex® Hi-Density TBE sample buffer (5x): 2 mL 5x TBE running buffer, 1.5 g Ficoll™ Type 400, 1 mL 1% bromophenol blue, 1 mL 1% xylene cyanol.
8. Glycerol: 20% (w/v) in water.
9. SYBR® Gold nucleic acid gel stain (1x): Dilute 5 μL of 10,000x stock into 50 mL of 1x Novex® TBE running buffer.
10. Safe Imager™ 2.0 blue light transilluminator.
11. SimplyBlue™ SafeStain.
12. PCR thermocycler instrument.

3 Methods

3.1 Bioinformatics
Bioinformatic analysis of macrodomains uses several programs for sequence alignment and secondary structure prediction. Beginning with an unknown protein or nucleotide sequence to be analyzed, similar sequences may be identified using a BLAST search [46]. If the sequence does not yield many hits, an iterative method such as
PSI-BLAST or PHI-BLAST [47] may be employed. Building on these results, a fold prediction algorithm may also be used to detect remote sequence similarities. FFAS (Fold and Function Assignment) incorporates this method [28]. FFAS performs protein sequence profile alignment to detect low levels of sequence similarity. All sequences used in and FFAS search must contain less than 1000 amino acids. Options are available for multiple sequences or for a pairwise sequence alignment. Several different databases may be searched, for example the Protein Data Bank (PDB) [48], Structural Classification of Proteins (SCOP) [49], Protein Families (Pfam) [50], Classification Of proteins in complete Genomes (COG) [51], and structural genomics sequence collections as well as complete genomes of individual organisms. Search results include a score that is derived from the FFAS algorithm. The algorithm arranges the matching profile alignments, with those that have the highest confidence on top. An FFAS score of $-9.5$ or lower is the criterion for a protein to match a known fold [52].

Jpred [29] provides secondary structure predictions about the protein sequence. Initially, Jpred uses PSI-BLAST to create multiple alignments, and the JNet neural network algorithm [53] is then used to make secondary structure predictions based on the multiple sequence alignment combined with predicted solvent accessibility. A one-letter sequence is used or a file with multiple sequences can be uploaded (plain text, FASTA, MSF, BLC, or Batch). The size limit per sequence is 800 residues in plain text format. The results include the secondary structure prediction confidence for each residue. Secondary structure components are marked with the letter H for $\alpha$-helix regions, E for $\beta$ strands and the letter B for buried residues. Each residue will also have a reliability score from 0 to 9, in which residues with a high score are predicted with higher confidence than those with lower scores [54]. Figure 1 shows an example secondary structure prediction for the putative conserved macrodomain in the bat coronavirus strain HKU4.

The likely locations of domain boundaries in proteins may be predicted by combining the analyses of protein sequences with these and other programs [55–57]. First, based on sequence alignments and conserved protein sequence motifs, putative macrodomains may be identified. Macrodomains share a globular structure of $\beta$-sheets flanked by $\alpha$-helices. While several human, bacterial, and viral macrodomains have been found to share a $\beta\alpha\beta\beta\alpha\alpha\beta\beta\beta\alpha\beta\alpha\beta$ architecture, many proteins in the macrodomain family contain additional $\alpha$-helices and/or $\beta$-sheets in their folds [58]. The arrangement of these predicted secondary structures assists in identifying this protein family. Slight variations, such as the addition of a single $\alpha$-helix at the beginning or end of the protein may occur [54, 59]. For example, in the SARS-CoV, the
SUD-N domain has a \( \beta\alpha\beta\alpha\beta\alpha\beta\alpha \) fold whereas the SUD-M domain has a \( \beta\alpha\beta\alpha\beta\alpha\beta\alpha\alpha \) fold [54]. In addition, conserved protein sequence motifs are associated with the macrodomain family. For example, the macrodomain ADP-ribose hydrolases employ the backbones of certain residues such as “GGGV” and “GIYG” to coordinate with the phosphate groups of ADP-ribose [60]. The MacroD1-like proteins employ a conserved aspartate and histidine as part of a catalytic triad to deprotonate a water molecule which can act as a nucleophile to attack the carbonyl carbon of an ADP-ribosylated protein [61, 62]. Moreover, additional motifs are

---

**Fig. 1** (a) Result of a FFAS database search using the bat CoV HKU4 nsp3 protein sequence. The top five results, their sequence identity, FFAS score and protein names are displayed. The Protein Data Bank served as the database for this search. The full result is truncated for legibility and virus names were added for clarity. (b) Jpred output from HKU4 query: \( \beta \)-sheet—E, \( \alpha \)-helix—H, buried amino acids—B. Confidence scores are listed at the bottom: (0) lowest, (9) highest. (c) Predicted macrodomain in HKU4. Predicted secondary structure, \( \beta\alpha\beta\alpha\beta\alpha\beta\alpha\beta\alpha\beta \), is shown above the sequence. Vertical lines and numbering indicate the start and endpoints of the predicted domain. Conserved residues of viral macrodomains that are directly involved in ADP-ribose recognition are boxed in black. Residues predicted to be involved in ADP-ribose recognition via output of the BioLiP database and COACH program are shown with green and black stars, respectively. Red stars indicate residues predicted to be involved in ADP-ribose recognition by both BioLiP and COACH. Black diamonds indicate residues that are conserved with respect to residues involved in interactions with the ligand, in the SARS-CoV macrodomain complex identified by Glide docking (Fig. 2a).
involved in the binding of ADP-ribose: the “NAAN” motif stabilizes the distal ribose through the sidechains of the second asparagine, the “SAGIF” stabilizes the two phosphate groups, and in the “DAIQ” motif, the carbonyl group of the aspartate forms hydrogen bonds with the adenine moiety of ADP-ribose in the binding pocket [23, 59, 63]. PAR glycosylases also contain a conserved “GGG” motif used to coordinate ADP-ribose but employ a “QEE” sequence where the two glutamate residues and a conserved N-terminal aspartate are responsible for trapping a water molecule for a nucleophilic attack on an ADP-ribose substrate for PAR cleavage [64, 65].

Putative macrodomains that lack conserved motifs may be identified by a fold prediction algorithm such as FFAS. This algorithm identifies remote fold similarities, and together with secondary structure prediction, may be applied to identify domain boundaries. Predicted secondary structure patterns from Jpred [29], PSIPRED [30], and other programs can identify disordered loop regions that often correspond to a domain boundary. To ensure only the domain of interest is chosen and disordered tails are not introduced into the construct, only a few residues beyond the predicted domain are incorporated into the construct of interest. Typically, hydrophilic residues are chosen as endpoints in a

---

Fig. 2 (a) Docked structure from Glide of the complex between ADP-ribose and the SARS-CoV macrodomain. ADP-ribose is shown as a stick model and colored by elements: white, hydrogen; green, carbon; red, oxygen; blue, nitrogen; purple, phosphorus. The protein is shown in a ribbon representation, with selected interacting residues in the binding cavity shown as ball-and-stick and labeled. For comparison, the X-ray crystallographic position of the ligand (PDB: 2FAV [59]) is shown as a wire-frame model in dark green. The two ligands exhibit a similar binding mode (RMSD = 0.74 Å). (b) Docked complex from ZDOCK between the SARS-CoV unique domain M (PDB: 2JZE [90]) and I14–Tel23, an antiparallel basket G-quadruplex from human telomeric DNA (PDB: 2KKA) [91]. The protein is shown as a ribbon model in green, and interacting residues of SUD-M are labeled and shown as stick models in red. The quadruplex is shown in cartoon mode in blue; interacting bases of the quadruplex are shown as cartoon ring mode in magenta. Predicted hydrogen bonds between the two molecules are shown as yellow dashed lines.
protein domain, such as a serine or aspartic acid. For many biochemical and biophysical studies, it is also helpful to reduce the number of cysteines and prolines; if these residues are near a domain boundary, and are not conserved and/or necessary for the protein’s activity. Cysteines oxidize to form disulfide bridges that can lead to protein oligomerization, while prolines introduce the possibility of conformational isomerization about the peptide bond. If cysteines are present in the sequence, reducing agents are introduced during purification. Prolines are not detectable in several common NMR experiments. At the N-terminus, destabilizing amino acids are avoided [66]. Disordered regions, domains, or proteins may also be predicted with software such as DISOPRED [31], SEG [32], or DisEMBL™ [33] that is specifically designed to identify them.

Bioinformatics analysis was applied to identify a potential new member of the macrodomain family in the bat coronavirus strain HKU4, which is phylogenetically close to the human MERS virus [67]. The sequence of the bat nonstructural protein 3, a large, multidomain protein, was analyzed by BLAST and FFAS. The sequence similarity to other proteins was searched by FFAS against the PDB database. As shown in Fig. 1, the sequence had strong similarity to known viral macrodomains. Secondary structure predictions from Jpred further showed a predicted secondary structure pattern consistent with that of a macrodomain. Multiple alignments identified conserved residues present in other macrodomains, such as the “DA” following the first β-sheet, the “GGGI” motif on the N-terminal end of α-helix 2, the “NAAN” that has the catalytically important second asparagine following β-sheet 2, and the “SAGIF” region N-terminal to α-helix 5. This was also supported by functional predictions, described in the next section. Using this information, we hypothesized that the bat coronavirus HKU4 nonstructural protein 3 (nsp3) contains a conserved macrodomain. To isolate this domain and study the protein further, we predicted domain boundaries. These boundaries are shown in Fig. 1. These boundaries are chosen a few residues outside the predicted secondary structures to maintain the predicted globular fold.

**3.2 Prediction of Function**

**3.2.1 Prediction of Protein Functional Classes**

Sequence-based or structure-based function prediction aims to generate a functional hypothesis about a protein before experimental data has been collected [68]. Online servers can be employed to predict protein structure and function, such as I-TASSER [36] and BindN [37]. I-TASSER is a program created by Yang Zhang’s group in 2008 to predict 3D protein structure from a given primary sequence. This prediction is useful when the protein structure is not known. I-TASSER will output a .pdb file that can be viewed via Pymol and Chimera [69], or similar molecular graphics programs. To start, all sequences must be in FASTA format and
have a length starting at ten residues and not exceeding 1500. Each submission will generate an ID that can be used to track the run’s progress. However, only one submission per IP address is allowed. I-TASSER generates a large ensemble of models and uses a clustering program, SPICKER [70], to select the most significant models based on pairwise structure similarities.

I-TASSER results are fed into the protein function database, BioLiP [45], and use a function prediction multi-server program, COACH [44], to further classify proteins. BioLiP is a database of protein–ligand interactions that has compiled structural information from the PDB database and data from published work. This database is used to provide a list of residues involved in potential ligand-binding sites that are responsible for catalytic activity. For uncharacterized proteins, the COACH algorithm was developed to predict ligand-binding sites and carry out functional analysis for I-TASSER. COACH analyzes sequences by incorporating comparative programs and ligand binding site prediction servers such as COFACTOR [38], FINDSITE [71], and ConCavity [72]. These programs work in a complementary fashion to predict ligand-binding sites for a queried protein and potential ligands that may interact with those given residues. A complete list of structural data from I-TASSER and functional data from COACH is provided in the output page provided in a link sent via email. COACH provides a list of the top five potential ligands and their binding residues based on their matches in the BioLiP database. Enzymatic profiles are ranked by the COACH algorithm based on the structural similarity of the queried protein to template proteins.

COFACTOR can also be used independently for novel protein sequences to produce information based on ligand binding sites and predicted functions [38]. COFACTOR uses both primary sequences in FASTA format and .pdb files as input for binding site and binding ligand predictions. Results from a COFACTOR prediction are focused on structural homology and three protein function types. First, a structural analysis against PDB structures is conducted, where the top ten best alignments are provided in the output. Next, protein functions are from three independent libraries: ligand binding prediction confidence (C), enzyme classification (EC), and gene ontology (GO). These identifiers are scored based on the confidence of their prediction [38]. These identifiers are linked to specific binding sites and protein functions from their respective databases. A description and related proteins with similar function are associated with each identifier.

Other online servers can be used for protein function prediction, such as ProFunc, SIFTER, and RaptorX. ProFunc uses PDB input files or identifiers to predict likely biochemical function [43]. The server will run a PDBsum analysis [73], a server that provides a brief structural overview, for every ProFunc submission. ProFunc
uses sequence and structural batch processes to compare against known protein structures [43]. ProFunc also uses 3D template searches of known protein ligand interactions to predict function [43]. Once ProFunc is complete, a color-coded output is produced. The output provides predictions on conserved motifs and folds, enzymatic activity, ligand binding site, and binding pocket similarities. Next, SIFTER (Statistical Inference of Function Through Evolutionary Relationships) is a server that makes predictions based on statistical methods [42]. Only a FASTA sequence is required for this tool. SIFTER will compile a list of proteins that match the queried sequence and organize the related protein function through a confidence score. The list of significant protein function matches is classified by a GO identifier which offers further explanation for each shared function. Lastly, RaptorX, a protein threading program, uses protein sequences to predict structure, binding sites, residue contact maps, solvent accessibility per residue, and structure properties through a multistep threading approach [74]. RaptorX constructs a structure template list from known proteins and optimizes the energy function to determine the quality of each predicted structure [39]. Using pairwise interaction preferences, RaptorX aligns the input sequence to the backbone of templates to output a final predicted structure [75]. The structural prediction method is employed by computational learning models that determine solvent accessibility, structured and disordered regions or the entire queried sequence and predicts enzymatic activity and ligands that may bind to the protein sequence [40]. To use RaptorX, only a name, email, and FASTA sequence is needed. Results from the structural prediction will provide a summary prediction of secondary structure, solvent accessible residues, and disordered regions. Results from the binding tool provide a list of ligands that are predicted to bind to the protein. Furthermore, the predicted contacts tool provides a contact map, top five predicted models in .pdb format, and an interactive model of the protein via a Jmol viewer.

The computational approaches described here are used to analyze the predicted domain in the bat CoV HKU4. Structural analysis predicted that the domain should be significantly similar to known macrodomains, such as the conserved macrodomains from the SARS-CoV [23, 59] and Middle East respiratory syndrome nonstructural protein 3 [76]. These results are shown in Table 3. The prediction of this protein as a macrodomain is consistent from both I-TASSER, ProFunc, and RaptorX analysis. Predicted functions included phosphate metabolic processes, hydrolase activity, and nucleic acid recognition, which suggests that this domain is involved with ADP-ribose and/or nucleic acid recognition by the consensus results from COACH, SIFTER, and ProFunc servers. BioLiP, COACH, and RaptorX provide consistent predictions that ADP-ribose is a likely to be a ligand for this domain. These servers
also predict that the conserved aspartate (D18), the second asparagine (N36) in the “NAAN” region, the residues G42-G43-G44, the S124-A125-G126-I128-F129 region, and the C-terminal asparagine (N153) are involved in the recognition of ADP-ribose. These motifs are also seen in other ADP-ribose binding macrodomains, in particular, SARS and MERS [23, 59, 76].

3.2.2 Prediction of DNA-Binding and RNA-Binding Proteins

BindN is a server that predicts DNA- and RNA-binding regions of a protein. This program was created by Liangjiang Wang and Susan Brown in 2006 [37]. Neural networks were trained to incorporate queried sequence information and determine residues that were solvent-accessible to provide reasonable DNA/RNA binding predictions. To increase the predictive capabilities of BindN, evolutionary data from a position-specific scoring matrix (PSSM) derived from a PSI-Blast search was introduced [77]. The most recent addition of support vector machines (SVMs), a set of learning algorithms, further increased the server’s sensitivity performance [37]. BindN only requires a primary sequence in FASTA format to run. Sensitivity and specificity parameters can be adjusted to increase the number of hits or reduce the number of false positives. The output will provide an assessment of binding propensity to DNA or RNA. A confidence value range from 0 (lowest) to 9 (highest) and prediction marker for binding residues (+) and non-binding residues (−) are indicated directly below each amino acid in the sequence. This program is useful in analyzing novel macrodomain sequences with potential nucleic acid-binding function [9, 78].

3.2.3 Computational Docking

Protein interactions with binding partners such as other proteins, peptides, nucleotides, and small molecules are the central focus of functional studies [79]. Computational docking has proven to be a valuable tool in the study of binding events, with the help of structural information from experiments and bioinformatics. There are two types of docking procedures in general: without any restraints (ab initio) and with restraints [80]. Restraints typically comprise experimentally derived information. Here we describe three different programs that were found to be useful in our laboratory, for different applications.

First, ZDOCK [81] is an ab initio docking program authored and maintained by Zhiping Weng and colleagues. The program searches all possible binding modes in the translational and rotational space between two molecules and evaluates each pose using an energy-based scoring function. The scoring function comprises interface atomic contact energy (IFACE) statistical potential, shape complementarity, and electrostatics [82].

1. Structure submission: Users can input the two structures to be docked on the initial submission page. Protein 1 is fixed during the docking, so it is better to use the larger structure as the
“protein,” and protein 2 as the “ligand,” or smaller structure. To input structures, one way is to specify PDB codes, if available, and the ZDOCK server can connect to the PDB database. The other way is to upload PDB files (see Note 4). After submitting the PDB files, the user should select individual chains if a multimeric protein is considered.

2. Selection of blocking/contacting residues. The user is asked to select residues that are involved in binding or to select residues that do not participate in binding interactions. This is an optional step to orient the ligand into possible binding sites. If experimental information is available regarding binding, it can be entered here. Upon selection, the residues can be highlighted in 3D visualization aided by Jmol [83].

3. The ZDOCK output includes the PDB top ten models generated by the program and preprocessed input PDB files. In addition, the results feature a Jmol visualization of the top-ranked docking models by energy score, and the center-of-mass positions of the top 500 ligands. There is also the capability to visualize or download any individual complex from the output. Alternatively, users can generate more sets of PDB files for predicted complexes by using a Java program or an executable file which are available from the ZDOCK website (see Note 5).

Second, HADDOCK [84, 85] (High Ambiguity Driven protein–protein DOCKing) is different from other ab initio docking methods such as ZDOCK because HADDOCK uses ambiguous interaction restraints (AIRs) to drive docking processes. AIRS can be obtained from information about the proteins, such as NMR chemical shift perturbation data, mutagenesis data, or any information regarding the interaction interface. HADDOCK requires a Linux or Unix environment with CNS (Crystallographic and NMR system) [86] for generating experimental restraints. A web server is implemented to provide an interface with different levels of control over the docking process, termed Easy, Expert and Guru. The Easy interface is freely accessible to nonprofit users, while Expert and Guru require upgrade access given by the program administrators.

Easy interface

1. Structure file preparation. Currently, HADDOCK uses CNS as its structure calculation engine, which requires a slightly different input file format compared to the PDB. Hence, file preparation is required if a PDB file is used as structure input. The differences between CNS and PDB format can be viewed in the HADDOCK manual (see Note 6).
2. Structure submission. Once structures are prepared, they can be entered in the HADDOCK Easy interface as the first and second molecule. The next important step is to define the active and passive residues (see Note 7) which are required to generate AIRs governing the docking process.

3. Viewing the results. Once docking finishes, a link containing parameters and predicted complexes will be sent to the Email provided by the user. HADDOCK uses its own scripts to pre-analyze the complexes. The link contains four top structures for each cluster that is pre-analyzed and listed by each clusters’ HADDOCK score. The HADDOCK score is a weighted sum score of different energy terms (electrostatic, van der Waals, desolvation and restraint violation), see Note 8.

Third, Glide (Grid-based Ligand Docking with Energetics) [87–89] is a high-throughput docking program that has been developed and updated by Schrödinger, LLC and now is incorporated into the Maestro suite with other programs required for small molecule–protein interaction studies. It is a conformational search-based program that employs a grid approximation of the nonbonded ligand-receptor interaction energy. A post-docking minimization is performed to return ten poses per ligand based on their energy score, distance and conformation. It provides extra precision (XP), standard precision (SP) and virtual high-throughput screening modes to trade off various needs for speed and accuracy.

1. Ligand and receptor preparation. Glide requires a preparation of receptor and ligand structures. The preparation is incorporated in the same Maestro suite as Glide. In the Maestro program, the user accesses the Protein Preparation Wizard and Ligand Preparation Wizard to perform this preparation. The goal is to check for missing atoms and connectivities from the NMR or X-ray crystallographic structures, assign the proper bond order, add hydrogens, and reduce multimeric proteins to a single unit. The ligand preparation will render two-dimensional chemical structures, such as ligand library files exported from molecule sketching programs (ChemDraw, Ghemical or PRODRG), to their three-dimensional forms with the preferred ionization states and correct stereochemistry (see Note 9).

2. Grid generation. The Glide program uses a grid map of the active site to search and dock ligands in an exhaustive manner [87–89]. Hence the grid map, which reflects the shape, electrostatic and other physical properties of the active site of the receptor, has to be ready before docking. The program provides a wizard called Receptor Grid Generation. The user employs a prepared receptor file from the Protein Preparation Wizard, then defines a Workspace box that will cover the active
site of the receptor during the docking. The other options include positional constraints that dictate the presence of certain functional groups in specific regions.

3. Docking. Setting up the Glide docking program involves selecting the grid file generated in step 2 and the ligand file or library prepared in step 1. The user then decides which docking mode to use and sets up the score penalties for conformational changes and torsions in docking the higher energy states of the ligand. The other important options to select are the extent of receptor flexibility during the docking and output formats for the resulting poses and scores.

4. Results analysis. Glide provides an interface to view the resulting poses and contacts between ligand and receptor. This viewing mode can be opened by, in the Project Table, the options Entry, View Poses, and Setup. The user then may analyze each docked pose in the Workspace, along with its Glide score and contacts. The export function can output docked structures to other types of files, such as .mol and .pdb. Glide scores, ionization states and other energy terms can be output to a spreadsheet.

The docking techniques described above can be used to recreate known binding interactions and predict new ones. For instance, Fig. 2a shows an example of a Glide re-docking result by using the crystal structure of the SARS-CoV macrodomain and its ligand ADP-ribose (PDB ID: 2FAV). All water molecules, including interacting ones, were removed and the protein was minimized with bound ligand in this example during the protein preparation. The interacting amino acid residues (D23, I24, N41, G47, G49, V50, S129, G131, I132, F133, N157), match experimentally determined ADP-binding motifs. In Fig. 2a, the D23 and I24 residues are shown to stabilize the adenine ring, the second asparagine (N41) from the “NAAN” motifs coordinates with the distal ribose, and both the “GGG” (G47-G48-G49) and the “SAGIF” (S129-G131-I132-F133) motifs interact with the two phosphate groups in ADP-ribose binding. The predicted binding mode of ADP-ribose matches the experimentally determined X-ray crystallographic position closely, with a root-mean-square deviation (RMSD) of heavy-atom positions of 0.74 Å (Fig. 2a).

In Fig. 2b, ZDOCK was used to investigate a macrodomain-nucleic acid interaction. The SARS-unique domain M is known as a G-quadruplex binding protein [54], but the exact binding model is still unknown. Figure 2b shows the ZDOCK-predicted complex between the SUD-M domain (PDB: 2JZE) [90] and I14-Tel23, an antiparallel basket G-quadruplex from human telomeric DNA (PDB: 2KKA) [91]. The resulting binding model shows predicted specific interactions between protein and nucleic acid in the conserved surface cavity of the
These interacting residues, N532, L533 (at the back side, not shown), I556, and V611, correlate closely with previous studies by NMR that showed significant chemical shift changes upon binding of G-quadruplex oligonucleotides [9, 90]. The predicted binding mode includes hydrogen bonds to the G-quadruplex phosphate backbone, and hydrophobic interactions between amino acid sidechains and nucleobases (Fig. 2b).

Figure 1 summarizes functional predictions for the HKU4 putative macrodomain. FFAS and Jpred predictions, used to identify domain boundaries and produce a construct for study, are shown. Additional data regarding functionally important residues are identified by stars, from functional prediction programs described above. In addition, we identify residues that are conserved between the bat HKU4 macrodomain, and residues involved in interactions with ADP-ribose in the SARS-CoV macrodomain from the docking run shown in Fig. 2a, described below. These residues make key interactions with the ADP-ribose ligand. Comparatively, most of these key residues are conserved, as shown in Fig. 1. For example, the aspartate (D18) after the first β-strand, the second asparagine after the second β-strand (N36), the G42-G43-G44 loop before the second α-helix, the S124-A125-G126-I128-F129 region after the fifth β-strand, and the C-terminal asparagine (N153) are all predicted to be involved in the binding of ADP-ribose. Structurally, these residues align with conserved ADP-ribose binding regions of both the SARS-CoV and MERS-CoV macrodomains. For example, D18 in the HKU4macrodomain aligns with D23 in the SARS-CoV macrodomain. The “GGG” motifs in both macrodomains, the G42-G43-G44 in HKU4 and the G47-G48-G49 in SARS, have an identical sequence alignment, further supporting the idea that the function of these proteins is similar.

Here, we describe methods for the bacterial overexpression and purification of viral macrodomain proteins.

1. Add 5 mL LB broth and 5 μL ampicillin (100 mg/mL) to six sterile culture tubes. Obtain transformed agar plate with BL21(DE3) cells colonies containing gene of interest. Pick one colony using a sterile plastic loop. Submerge the tip of the loop in the broth and gently shake (see Note 10). Place tube at an angle in an incubator shaker at 37 °C for 12–18 h at 260 rpm. Using six 500 mL culture flasks, add 100 mL LB broth to each and autoclave for 1 h 15 min. Cool the flasks until cool enough to touch with the bare hand and add 100 μL ampicillin (100 mg/mL) to each (see Note 11).

2. Warm 100 mL LB broth in a 500 mL flask to 37 °C. Using a sterile pipet, transfer each 5 mL overnight culture to the fresh
100 mL of broth. Allow culture to grow for 12–18 h. Add 700 mL LB broth to each of six 2.8 L culture flasks and autoclave for 1 h 15 min.

3. Cool the flasks until cool enough to touch with the bare hand. Add 700 μL of ampicillin (100 mg/mL) and parafilm the top closed. Cool on the bench.

4. After 12 h use a sterile pipet to dilute the 100 mL culture to 700 mL. Allow the culture to grow to an optical density at 600 nm (OD₆₀₀) of 0.5 and reduce the temperature to 18 °C. Once the OD₆₀₀ of the culture reaches 0.6, add 805 μL of 1 M IPTG to each large flask. Allow culture to grow for 12–18 h. Warm broth before adding culture.

5. Using centrifuge tubes, centrifuge the culture at 5000 × g for 5 min. Discard the supernatant into a container and treat with 50% (v/v) bleach. Store collected pellet in −80 °C freezer.

1. Inoculate one colony into 5 mL LB broth with ampicillin. Place tube in an incubator shaker at 37 °C for 12–18 h.

2. To a 500 mL flask add 75 mL of 1× M9 solution. To each of four 2 L flasks add 241 mL of 1× M9 solution. Autoclave. Prepare and filter minimal media mix (Table 2). Store at 4 °C until use.

3. Add 2.5 mL of minimal media mix to the 500 mL culture flask. Add 9 mL of minimal media mix to each of the 2 L flasks. Foil and parafilm flasks closed.

4. Place 500 mL flask in the shaker to warm before adding culture. Allow culture to grow at 37 °C for 12–18 h.

5. Allow 2 L flasks to warm to 37 °C before adding saturated overnight culture. Transfer 12.5 mL of culture to each 2 L flask and allow culture to reach an OD₆₀₀ of 0.5. Reduce temperature to 18 °C. Induce culture at an OD of 0.6–0.8 by adding 250 μL of 1 M IPTG. Allow the culture to grow for 12–18 h. Alternatively, the temperature may be maintained at 37 °C and the cells grown for an additional 2–3 h.

Table 2
Reagents for minimal media mix (step 2, Subheading “Expression in Minimal Media”)

| Reagent                   | Volume |
|---------------------------|--------|
| 1 M MgSO₄                 | 2 mL   |
| BME vitamin mix           | 10 mL  |
| Solution Q                | 2 mL   |
| 20% Glucose               | 20 mL  |
| NH₄Cl or ¹⁵NH₄Cl          | 1 g    |
Table 3
Selected results from protein functional analysis of a putative macrodomain in the bat HKU4 coronavirus nonstructural protein 3. **Top section:** Comparison of domain prediction by I-TASSER structural alignment algorithm and ProFunc structural scan. I-TASSER provides a list of PDB templates and structural motifs that align to a novel sequence. The list provides the top three matches. ProFunc scans databases for structures that have similar structural motifs. **Middle section:** Predicted functions of the HKU4 domain from COACH, SIFTER, and ProFunc servers based on their Gene Ontology (GO) descriptions. **Lower section:** Predicted ligands from the BioLiP database, ProFunc and RaptorX servers are listed next to aligned PDB structures.

| I-TASSER/ProFunc | Identifier | Structural motif |
|------------------|------------|------------------|
| I-TASSER alignment | PDB ID: 2ACF | Macrodomain |
| | PDB ID: 5DUS | Macrodomain |
| | PDB ID: 3EWQ | Macrodomain |
| ProFunc | ProSite: PS51154 | Macrodomain |
| | Pfam: PF01661 | Macrodomain |
| | SMART: SM00506 | Macrodomain |

**COACH/COFACTOR**

| Molecular function | Biological process | Cellular component |
|--------------------|--------------------|--------------------|
| Molecular function | Biological process | Cellular component |

**Predicted ligands**

| BioLiP/ProFunc/RaptorX | Predicted ligands | Structurally similar proteins |
|------------------------|------------------|-------------------------------|
| BioLiP | Adenosine-5-diphosphoribose (APR) | Venezuelan equine encephalitis virus (PDB ID: 3GQO) |
| | Glycerol | Feline coronavirus (PDB ID: 3EW5) |
| | Nucleic acid | Chikungunya nsp3 virus macrodomain (PDB ID: 4TU0) |

(continued)
after induction; or intermediate values may be used. Conditions should be chosen based on the expression properties of the desired protein [92].

6. Centrifuge culture at 5000 × g for 5 min. Discard supernatant. Store pellet at −80 °C.

3.3.2 Protein Purification

The protocol below is a purification description for predicted macrodomain proteins of the bat coronavirus HKU4.

1. Thaw pellets on ice with lysis buffer (10 mL per gram of pellet).
2. Resuspend pellet by slowly adding lysis buffer to the container and gently mixing the solution with a clean glass stir rod. The lysis solution should be homogeneous before sonication.
3. To sonicate suspension, insert tip ¼ in above bottom of beaker and sonicate for a total time of 3 min (see Note 12). Pulse for 15 s after a 45 s pause at a power of approximately 42 W.
4. To separate lysate from cell debris, centrifuge at 20,000 × g for 20 min. Decant the supernatant into a clean container. Use a sterile syringe and a 0.22 μm low protein-binding syringe filter to filter the entire lysate volume (see Note 13). Take a small sample of the cell debris pellet and lysate sample for later SDS-PAGE gel analysis.

Column Chromatography

The protocol described is based on affinity purification by nickel affinity chromatography followed by cleavage with tobacco etch virus protease, a second nickel affinity chromatography step and a final size-exclusion chromatography step. In this system we make use of the 6xHis tag that was expressed N-terminal to the protein which binds with high affinity to Ni²⁺. The affinity difference of the expressed protein versus cellular proteins allows for a highly selec-
Active separation. Another advantage of this purification method is its speed. Using the ÄKTA™ purification system, we can quickly remove a large amount of impurities with ease. Below is a step-by-step description of nickel affinity purification using the ÄKTA™ purifier system. Each step was conducted at 4 °C.

1. Equilibrate the purification lines by executing “PumpWashBasic” in the Pump dialog box for both Line A and Line B in water. Wash column, detector, lines, and system with filtered, degassed water so that no ethanol remains in the system. Equilibrate the column with 5 column volumes (CVs) of running buffer. Execute “AutoZeroUV” in the “Alarms and Monitors” dialog box to set a baseline value of zero UV absorbance. Set a pressure maximum for the column of 0.3 MPa.

2. To load sample on column, the P-960 sample pump collection tube is placed in running buffer and equilibrated with a 15 mL volume. Place the P-960 collection tube in the sample and start a flow to the column with a 1 mL/min flow rate. Increase flow rate to 5 mL/min in 1 mL/min increments. Fraction sizes of 3–10 mL are collected for SDS-PAGE analysis (see Note 14).

3. After all of the sample has been loaded, the P-960 is turned off and an additional 10 CV of running buffer is passed over the column to wash away any unbound impurities from the column.

4. The desired protein is eluted from the column using a shallow gradient from 100% running buffer to 100% elution buffer. The gradient should reach 100% elution buffer in a minimum of 10 CV for optimal resolution.

5. Run 10 CV of elution buffer to ensure that the entire protein sample has eluted from the column.

6. Wash column with 5 CV of running buffer followed by 10 CV of filtered, degassed water to regenerate the column. Column can be reused for another purification, or, if this is followed with 20% ethanol, stored.

Using tobacco etch virus protease, the 6xHis tag is cleaved from the protein, decreasing its affinity to the Ni²⁺ column. The imidazole is dialyzed out of the protein sample prior to overnight cleavage to optimize protease activity.

1. Pool fractions that contain expressed protein, as assessed by SDS-PAGE analysis.

2. Prepare 6 L of Ni²⁺ affinity running buffer to dialyze protein sample.

3. Use Snakeskin® (3500 MWCO) to dialyze pooled fractions. At 4 °C, exchange dialysis buffer every 4 h.

4. Transfer sample to a Falcon tube and add 80 μL AcTEV™ protease (800 units) at room temperature overnight. After cleavage reaction, filter sample with syringe filter.
1. Repeat Ni$^{2+}$ column purification steps. Protein should not bind to column, instead elute in the flow-through due to the removal of the 6xHis tag.

2. Repeat SDS-PAGE analysis and pool the protein-containing fractions.

3. Concentrate sample with VIVASPIN™ 20 concentrator at $6000 \times g$ for 20 min intervals. Pipet sample to mix between spin intervals. Concentrate to a final volume of 13 mL or less (see Note 15).

SEC is the final step in this purification scheme. This chromatographic approach is a popular finishing step due to its ability to separate protein samples based on their molecular weight. Inactive protein aggregates are separated and removed [93].

1. Attach 50-mL Superloop™ to ÄKTA™ purifier and load the concentrated protein sample into sample chamber. To load sample, allow the main pump (Line A) to fill the superloop with buffer. After filled, turn the injector valve to “INJECT” and inject sample with a plastic syringe into the inject valve port. Once all of the sample has been added, turn the injector valve to “LOAD”. Load the sample to the HiLoad™ 26/600 Superdex™ column at 1 ml/min. Increase flow rate to 2.6 mL/min without exceeding 0.3 MPa. Flow 1.5 CV of running buffer through the column. Collect fractions.

2. Repeat SDS-PAGE to identify fractions with purified protein. Pool fractions.

3. Using a new concentrator, add ultra-pure water and spin in concentrator for 10 min to rinse membrane preservatives from membrane 3 times. Concentrate pooled fractions until total volume is 500 μL. Spin at a maximum of $6000 \times g$ for 20 min intervals. Use pipette to mix retentate after each spin. Continue centrifugation until sample volume is 600 μL.

As a final step of purification, recombinant overexpressed HKU4 domain purification was completed via size-exclusion chromatography. Because SEC is based on the molecular size of a protein, the elution time of a protein can be calibrated to determine its molecular weight (smaller proteins have a longer path and elute later while larger proteins have a shorter path and elute earlier) [93]. The estimated elution volume of an 18 kDa protein is 245 mL based on molecular weight standards, and the resolved peak at 250 mL is the purified protein. In Fig. 3, the molecular weight of the protein was confirmed by SDS-PAGE analysis and also showed the protein sample did not contain detectable impurities. The secondary peak at 330 mL is indicative of low MW impurities that were present in the protein sample prior to SEC; these were too low in concentration to appear on the gel.
Nuclear magnetic resonance (NMR) spectroscopy takes advantage of the characteristic nuclear spin angular momentum of atomic nuclei to probe the behavior of molecules through interactions with a radiofrequency field. For biomolecular studies, the nuclei of greatest interest are $^1\text{H}$, $^{13}\text{C}$, $^{15}\text{N}$, $^{19}\text{F}$, and $^{31}\text{P}$ [94]. Each nucleus resonates at a characteristic frequency that changes when introduced to different local microenvironments. The observed changes provide valuable information about the molecular structure and conformation of a given sample [95]. For example, one can use the information collected from $^1\text{H}$, $^{15}\text{N}$, $^{13}\text{C}$ and $^2\text{H}$-based experiments to gain information on high-resolution structure, protein stability, dynamics, folding pathways, enzyme mechanisms, and protein complex assembly, in both the solution and solid state [96–100]. The methods described below provide guidance to conducting and analyzing NMR techniques for protein–ligand interactions by saturation transfer difference (STD) NMR and by chemical shift mapping.

**NMR Spectroscopy**

3.3.3 **NMR Spectroscopy**

NMR Sample Preparation

1. Exchange sample into NMR buffer via ultrafiltration (step 3 from Subheading “Size-exclusion Chromatography”).
2. Add 3% D$_2$O, 5 mM DTT-d$_{10}$, and 0.02% NaN$_3$. 

![Chromatogram of the HKU4 putative macrodomain size-exclusion chromatography purification.](Fig. 3a)

The protein elution is indicated by the major peak. UV detection at 254 nm and 280 nm is indicated by the red and blue traces. 

![Elution of protein (18.4 kDa) monitored by SDS-PAGE (4–12% Bis-Tris gel) of selected fractions collected from the chromatography run in (a).](Fig. 3b)
1. Prepare 100× stock ligand concentration in 10 mL NMR buffer (10 mM sodium phosphate, pH 7.2, 150 mM NaCl).

2. In a 1.5 mL Eppendorf tube, prepare 600 μL NMR sample. Final concentrations: protein (20 μM), ligand (2 mM), 3% D$_2$O, 5 mM DTT-d$_{10}$, and 0.02% sodium azide.

3. Obtain a Kimble NMR tube and use extended tip pipette to wash with ultra-pure water. Rinse the tube three times with buffer before placing the sample in the tube. Ensure no bubbles are present in the NMR tube.

The STD-NMR experiment is a commonly used ligand screening technique for rapid analysis of protein–ligand interactions. This method is based on the transfer of magnetization from protein to ligand [96]. This process occurs by the application of a low-power radiofrequency pulse to the protein [96]. This saturation builds up through the entire protein through intramolecular 1H-1H nuclear Overhauser effect transfer. This saturation will also transfer to other protons located close to the protein in a binding pocket [101]. Thus, ligands that bind to a selectively irradiated protein will also become saturated. Experimentally, this process is useful because the saturated protons from a bound ligand dissociate, where they relax more slowly than in the bound state [101]. The saturation may be observed in a difference spectrum. The spectrum with on-resonance saturation is subtracted from the spectrum with off-resonance saturation [102]. The difference spectrum contains only signals from the ligand involved in the binding event. Consequently, this method has become widely used to identify compounds that bind to proteins of interest, and to carry out epitope mapping of binding compounds.

Notably, there are a few drawbacks. If a ligand binds too weakly then it will not stay bound long enough to be saturated sufficiently, and conversely, if it binds too tightly then the saturated ligand will relax too quickly to be detected [96]. For a good STD-NMR experiment, the ligand should have dissociation constants, $K_D$, ranging from approximately $10^{-3}$ to $10^{-8}$ M. Therefore, this method is quite robust in its ease of detecting protein–ligand interactions.

The method below describes an STD-NMR analysis with a putative macrodomain from the bat HKU4 CoV and ADP-ribose, using Bruker instrumentation.

1. Start the Topspin 3.5 program from the computer controlling the spectrometer.

2. Set the temperature to 298K, or preferred temperature. (This temperature setting is sample-dependent.)

3. Place sample in the spinner. Clean the NMR tube of finger-prings and dust with a Kim-wipe. Adjust the position of the tube with the depth gauge.
4. Place the tube and spinner in the sample carousel under the sample collecting tube. Select “ij” to inject the sample. Once the sample indicator light on the spectrometer is a steady yellow, activate the field-frequency lock by typing “lock” and selecting the appropriate solvent, here 90% H₂O/10% D₂O. Open the lock window to view the locking process.

5. Open a new 1D dataset and type “atmm” to tune and match the probe. Align the wobble curve so the minimum rests at the bottom. Save position and exit.

6. Type in command “topshim” to shim using the gradient shimming routine. One-dimensional or three-dimensional gradient shimming may be used.

7. Type “rpar” and find a set of baseline experimental parameters to begin with. It is usually useful to begin with a simple one-dimensional NMR experiment, e.g., zgpr or zgesgp.

8. Calibration: Execute “pulsecal” and obtain the pulse length for the 90° excitation pulse. Adjust the transmitter frequency via the “gs” window and offset tab until the free induction decay (FID) area is at a minimum. This should be done in a water suppression experiment. Execute “getprosol 1H (hard pulse length or P1) “PLW1 (in W)”. This will calibrate the power levels for the entire experiment.

9. Set the number of scans to 16. Execute “zg”.

10. Examine the spectrum and identify regions where only protein signals are found. These regions should be at least 1 ppm away from ligand signals. These regions are appropriate for selective saturation of the protein.

11. In a second dataset, read in a baseline parameter set for STD-NMR, such as STDDIFFESGP.

12. Edit the file “FQLIST,” which defines transmitter positions to be used in the pulse sequence, including for selective saturation of the protein. For only one irradiation point, enter the frequency (in Hz) from the 1D spectrum. A sample list is shown in Table 4.

13. For a quick scan to test the pulse sequence, set NS = 16. For a full scan, set NS = 512. Execute “zg”.

14. Processing: After acquisition is complete, type “stdsplit”. Select the first spectrum and execute “efp”. This will process

| Table 4 |
|-----------------|-----------------|
| **Frequency list in Bruker TopSpin format for STD-NMR setup** |
| 1 | 0700.32 |
| 2 | 268.63 |
| 3 | −50,000 |
the time-domain data and produce the reference spectrum. This will be identical to the first 1D spectrum from step 10. Select the second spectrum, which is the STD-NMR difference spectrum, and execute “efp”. This will produce any $^1$H signals from a ligand that binds to the protein during the saturation period.

The STD-NMR difference spectrum will yield only signals from ligand hydrogens that are involved in interactions with the protein [102]. Based on the distinct chemical shifts of each $^1$H, information on how the ligand binds to the receptor can be determined.

The protocol discussed in this section was used to detect ligand binding to the HKU4 putative macrodomain by showing enhancement of the NMR signals from ADP-ribose. Further analysis using this protocol can be used to map the ligand in the binding pocket and determine a binding constant ($K_D$). The binding constant of the HKU4 macrodomain is expected to be similar to those of other viral macrodomains, such as that of the SARS-CoV ($\sim$24 μM) [59].

The local chemical microenvironments of a protein are sensitive to changes in solvent and molecular interactions. This phenomenon is useful when mapping structural changes and classifying binding interactions of a protein ligand complex. The method described here measures the chemical shifts of a protein, via a $^{15}$N, $^1$H]-HSQC spectrum, at different ligand concentrations. Based on the relative movement of peaks from each titration step, a map of the binding pocket can be determined [96].

**Control**

1. In an Eppendorf tube, prepare 600 μL sample. Final concentration: protein (50 μM), 3% D$_2$O, 5 mM DTT-d$_{10}$, and 0.02% NaN$_3$.
2. Repeat steps 1 through 9 described in the previous section.
3. Execute “iexpno” and “rpar”. Read in a baseline parameter set for a $^{15}$N, $^1$H] correlation experiment, or HSQC, such as “HSQCETFPF3GPMI”. Select the keep parameters option.
4. Select “atmm”. Tune and match the $^{15}$N channel and the $^1$H channel. Save and close.
5. Type in command “topshim” to shim using the gradient shimming routine. One-dimensional or three-dimensional gradient shimming may be used.
6. Execute “getprosol 1H (hard pulse length or P1) “PLW1 (in W)”.
7. Set sweep width (SW): F2 = 14 ppm and F1 = 40 ppm.
8. Set NS = 16. Execute “zg”.

---

**Chemical Shift Perturbation (CSP)**
Titration

9. Prepare 1000× stock concentration of ligand in 15 mL of NMR buffer (10 mM sodium phosphate pH = 7.2, 150 mM NaCl).

10. Add enough ADP-ribose for a final 0.2:1 ADP-ribose: protein molar ratio in a new Eppendorf tube. Remove the entire sample volume from the NMR tube and place into the Eppendorf tube. Gently mix and return to NMR tube.

11. Repeat step 10 for each titration point. Add enough ADP-ribose to increase the molar ratio of the sample in each experiment as follows: 0.5:1, 1:1, 10:1.

12. Analyze the spectra by selecting the control dataset and dataset with added ligand for the other HSQC spectra at different concentrations.

The control spectrum is the baseline spectrum in the absence of ligand, used to determine the chemical shifts of the apoprotein. Once ligand is added to the protein sample, changes in the spectrum due to ligand binding may be observed [103]. In a \(^{15}\text{N},^{1}\text{H}\)-HSQC, each signal results from a covalently bonded \(^{15}\text{N},^{1}\text{H}\) pair. With a few exceptions, the correlation peaks represent each amino acid in the protein sequence and are used to map the binding pocket of the protein. However, for unambiguous characterization of the binding pocket, assignment of the backbone is necessary [96].

Perturbation of chemical shifts occurs under three general conditions: fast, slow and intermediate exchange rates. Fast exchange indicates that the off-rate \((k_{off})\) for protein–ligand interaction is much greater than the frequency difference (chemical shift difference) in Hz between the bound and free states of a given peak. This condition is often seen for weakly binding ligands, with binding constants \((K_D)\) in the high \(\mu\text{M}\) to \(\text{mM}\) range [96]. In a titration with increasing ligand amount, peaks in the HSQC spectrum are observed to gradually shift from their original position toward the chemical shift characteristic of the bound state. In fast exchange conditions, the observed chemical shift is a weighted average of the shifts in the free and bound states [104]. \(K_D\) may be determined by a fitting procedure. Slow exchange occurs when the off-rate is much less than the chemical shift difference between these two states, a condition that is often satisfied for ligands that bind with sub-micro-molar affinity. In this condition, two separate signals for the bound and free states of a given peak are observed. Intermediate exchange is observed when the off-rate is comparable to the chemical shift difference and is characterized by peak broadening together with chemical shift changes [96, 104].

The CSP method is useful in screening potential ligands for novel proteins. Testing multiple ligands and mapping their
interactions with the protein can help shed light on proteins with unknown function. In particular, predicted macrodomains have characteristic functions that can aid in their classification, such as ADP-ribose recognition. Furthermore, CSP aids in the analysis of protein–ligand binding affinity and offers structural information characteristic of a protein–ligand interface.

The steps in this protocol describe a biochemical assay that is used to identify protein–nucleic acid complexes. We describe a protocol that is used for a variety of protein and ligand samples under native conditions.

Preparation of Nucleic Acids

1. Add nuclease-free water to lyophilized stocks to make 5 mM stock concentration. Gently mix to dissolve (see Note 16).
2. Dilute stock oligonucleotide to a working concentration of 40 μM in EMSA buffer. For guanine quadruplex sequences, place samples in thermocycler and set annealing protocol to increase temperature to 95 °C for 5 min, then cool stepwise (28 steps) to 4 °C over 12 h.
3. Add corresponding buffer volumes to each tube as described in Table 5.
4. Add corresponding protein samples from stock protein. (see Note 17)
5. Add 5 μL of annealed DNA to each sample to yield 5 μM total DNA concentration.
6. Allow sample to equilibrate at 25 °C for 1 h.
7. Add 5 μL of 20% glycerol to each sample to a final concentration of 3%. This helps to stabilize the protein–ligand complex for gel loading.
8. Add 5 μL of TBE sample buffer to each sample.
9. Assemble TBE native PAGE gel (12 wells, 1-mm thick) in the electrophoresis system and lock into place. Add 1× TBE running buffer to middle and outer chamber.

Table 5
Specified protein and buffer sample volumes for electrophoretic mobility shift assay (EMSA). Volumes were chosen to increase protein concentration by 20% in each step

| Sample no. | 1   | 2   | 3   | 4   | 5   |
|-----------|-----|-----|-----|-----|-----|
| Protein   | 0 μL| 1.7 μL| 4.4 μL| 8.9 μL| 15 μL|
| Buffer    | 15 μL| 13.3 μL| 10.6 μL| 6 μL| 0 μL|
10. Load 15 μL of each sample.
11. Run gel at 100 V for 45 min at 4 °C.
12. Prepare 50 mL of 1× SYBR® Gold stain. Add stain to plastic container and cover with aluminum foil to prevent exposure to light. Shake for 40 min. View with a Safe Imager™ transilluminator and record the gel image with a camera. Dispose based on local regulations.
13. Place the gel in a new plastic container. Add 100 mL of water. Close container, microwave for 1 min and shake for 1 min. Decant water and repeat twice.
14. Add 20 mL SimplyBlue™ SafeStain to cover the gel. Close and microwave for 1 min. Place container on shaker for 5 min. Decant stain and add water. Allow the container to wash stain from gel on shaker for 10 min before viewing. Record gel image with a camera.

This method is useful when determining if the protein is a receptor for a given oligonucleotide. If so, the protein–ligand complex will appear as discrete bands at higher molecular weight, whereas free oligonucleotides will travel farther through the gel due to their lower molecular weight. The increased molecular weight of the complex formed will prevent the complex from sieving through the gel as fast as the free oligonucleotide [105]. The gels are stained with protein stain to confirm the colocalization of the oligonucleotide and protein.

EMSA has been used to identify nucleic acid binding partners for macrodomains [9]. Assays based on varying DNA and RNA sequences were used to determine sequence specificity preferences and estimate binding affinities of novel viral macrodomains. Analysis of structured oligonucleotides was conducted to compare size and sequence preferences of the HKU4 macrodomain. Preliminary studies suggest that nucleic acid binding may be affected by sequences surrounding the core domain (data not shown).

3.3.5 Circular Dichroism Spectroscopy

1. Open nitrogen flow valve and turn on CD spectrometer. Open Spectrum Manager software. Allow nitrogen to flow through the optics for 10 min.
2. Select spectrum measurement option. Make sure that all pre-check parameters display “OK” before proceeding with the instrument.
3. Add 200 μL of solvent to the 0.1 mm quartz cuvette (see Note 18). Use the sample holder to verify the volume is higher than the light source opening. Place cuvette into the sample chamber with the sample block. Return the sample chamber cap. Select experimental parameters for a CD scan, including: wavelength range (260 nm—start, 190 nm—end); response (4 s); accumulation (2 scans/experiment). Solvent spectrum should display a line with no or very little slope.
4. Remove solvent from cuvette and add 200 μL of the protein sample. Run CD scan.

5. Subtract baseline solvent spectrum from protein sample. Smooth curve for final output.

6. Convert data to mean residue ellipticity, \( \theta_{\text{MRE}} \), using a spreadsheet program. Run SELCON3 or other program for deconvolution of CD data. Output will include global secondary structure and random coil content of the protein.

CD spectroscopy is a useful technique in determining the secondary structure of a protein. Common features of \( \alpha \)-helix (minima at 222 nm and 208 nm; maximum at 190 nm), \( \beta \)-sheet, (minimum at 218 nm; maximum at 196 nm), and random coil (minimum at 195 nm; maximum at 212 nm) can be seen from the CD spectra. The spectra can be deconvoluted to determine the global secondary structure content of a protein, using various methods and software (recently reviewed by N.J. Greenfield [106]). The assumption used by all methods is that the observed spectrum is a linear combination of the spectra of its secondary structural components plus a noise contribution from aromatic groups and prosthetic groups. For example, a CD measurement that has strong minima at 222 nm and 208 nm is expected to have a large fraction of \( \alpha \)-helix structure. This is helpful when verifying the fold type and stability of a protein sample. These analyses are implemented by software such as SELCON3 [107] which takes mean residue ellipticity, or \( \theta_{\text{MRE}} \), as input. SELCON3 will analyze these data and report the content of each secondary structure in the analyte.

The combination of bioinformatic approaches, protein function prediction and computational docking has allowed the identification of new members of the macrodomain family based on protein sequence alone. These sequences have also been used to develop homology models. To validate these predictions, a system of biochemical and biophysical approaches for characterizing protein structure and ligand interactions may be employed. These methods are useful for the study of proteins with no previously known structural or functional data, as described here for a putative macrodomain in the bat coronavirus HKU4.

## 4 Notes

1. FFAS is also available for download.

2. The glucose solution is best prepared fresh to minimize the chance of microbial contamination.

3. The pH of the buffer should be no more than 1 pH unit away from the theoretical pI of the protein. Use 20 mM Tris for more basic buffers. Use 10 mL of lysis buffer for every gram of cell paste recovered from culture growth.
4. The ZDOCK server will input the first state of PDB files containing multiple states (e.g., NMR ensemble). The manual submission mode will input all states but treat atom positions collectively; hence, it is better to separate states before manual submission (e.g., split_states command in Pymol).

5. The ZDOCK output file is given the extension .out. An explanation of the output file is given on the ZDOCK server website (http://zdock.umassmed.edu/zdock_output_format.php). This file also contains the displacement of the final structures relative to the initial structures and their corresponding ZDOCK scores. To generate all the predicted complexes, the ZDOCK program will need the initial structure files, also in the result link, and this output file to calculate the relative displacement and generate individual PDB files.

6. The Python scripts for converting and preparing proteins in PDB format are provided in the GitHub repository of HADDOCK, where it also explains how HADDOCK deals with additional PDB structural information such as missing density, cofactors, and alternative positions. For DNA/RNA structures in PDB format, the one-letter code (residue name, column 18–20 in PDB format) of the PDB file needs to be converted to a three-letter code in the CNS PDB format (see Table 6). This should be easy to accomplish by a search and replace function in Microsoft Word or other text replace scripts.

If one wishes to dock a DNA/RNA sequence with no solved structures, the 3D-DART server (3DNA-Driven DNA Analysis and Rebuilding Tool) [108] is provided to generate custom 3D structural models of DNA and RNA with control over the local and global conformation such as nucleic acid type (A-form or B-form), bend-angle, location of the bend-angle, and custom values for base-pair step parameters.

7. The “active” residues are those that have been experimentally determined to be involved in the interaction between the two molecules, and also solvent-accessible (either main chain or side chain relative accessibility should be >40–50%). This accessibility cutoff is not a hard limit and should be adjusted by the user based on their understanding about the interface. The “passive” residues are solvent-accessible surface neighbors of active residues.

8. All results are saved in structures/it1/analysis and structures/it1/water/analysis. These statistical and energetic analyses are described in the Analysis part of the online HADDOCK manual.
9. The detailed tutorials and user manuals for these preparations are available at Schrödinger’s website. In the protein preparation, there are three tabs: Import and Process are to perform basic tasks and fix problems from crystal structures such as missing density and bond orders. The Review and Modify tab is to delete unwanted chains (e.g., for multimeric proteins) and water molecules. The Refine command allows the user to perform energy minimization on the structures by changing hydrogen bond orientations and side chain orientations. The automatic restrained refinement is done using an OPLS2005 force field with heavy atoms convergent to a stated RMSD, which by default is 0.3 Å. If the user sets this RMSD to zero, then all the heavy atoms will be fixed during the minimization, with only hydrogens being minimized. In the ligand preparation, the program can take one or a library of 2D chemical molecules and assign hydrogens, ionization properties, stereochemistry and generate tautomers.

10. Inoculate near an open flame.

11. If necessary, the broth with antibiotic added may be stored at 4 °C overnight. The flask should be covered with foil and foil sealed with Parafilm.

12. To keep the resuspended culture cool during sonication, transfer resuspended cells to a glass beaker in an ice bath.

13. When filtering the lysate in preparation for HPLC chromatography, be careful not to make bubbles which can precipitate protein in the sample. Change syringe filters if it becomes clogged and filtering becomes manually difficult.

14. Set the flow path of the ÄKTA™ purifier to flow through the column and out the fraction collection valve. This equilibrates all of the flow lines in your system and ensures proper flow and pressure for purification. Load a maximum of 20 times the column volume as we have seen this will overload the binding capacity of the column.

15. 13 mL is the load volume limit for size-exclusion chromatography because a higher sample volume would compromise the separation efficiency of the purification step. Particles separate based on their size through the pores of the column. If the sample volume is too large, then the pores can overload and

**Table 6**
The conversion of one-letter code in PDB format to the three-letter code in CNS PDB format

| One-letter code in PDB | A   | C   | U   | G   | DA  | DC  | DT  | DG  |
|------------------------|-----|-----|-----|-----|-----|-----|-----|-----|
| Three-letter code in CNS | ADE | CYT | URI | GUA | ADE | CYT | THY | GUA |
the accessible surface area of the column will decrease. This will prevent smaller particles from taking a longer path through the column and decrease the resolution of the purification. To achieve the best resolution, the sample volume should be 5% of the entire column volume or less.

16. If using RNA, handle stock solutions in a dedicated RNA area and pipette set to reduce exposure to nucleases that are often present in laboratories.

17. Protein concentration is increased by 20% in each sample to provide a reasonable range of concentrations.

18. Blank run. Recommended protein concentration: <10 μM. Recommended salt concentration: <100 μM.

**Acknowledgments**

This work was supported by University of Alabama at Birmingham Faculty Startup Funding, NIH NIGMS GM119456-01, the University of Alabama at Birmingham Department of Chemistry, and the NSF Bridge to Doctorate Program. We thank Jeffrey McDonald and Sadanandan Velu for assistance with the Glide program. We thank members of the Johnson laboratory for helpful discussions and technical assistance. The UAB Cancer Center NMR facility is supported by a CCSG Grant P30 CA-13148 from the National Cancer Institute.

**References**

1. de Groot RJ, Baker SC, Baric RS, Brown CS, Drosten C, Enjuanes L, Fouchier RA, Galiano M, Gorbalenya AE, Memish ZA, Perlman S, Poon LL, Snijder EJ, Stephens GM, Woo PC, Zaki AM, Zambon M, Ziebuhr J (2013) Middle East respiratory syndrome coronavirus (MERS-CoV): announcement of the Coronavirus Study Group. J Virol 87(14):7790–7792. doi:10.1128/JVI.01244-13

2. Zumla A, Hui DS, Perlman S (2015) Middle East respiratory syndrome. Lancet 386(9997):995–1007. doi:10.1016/s0140-6736(15)60454-8

3. Peiris JS, Guan Y, Yuen KY (2004) Severe acute respiratory syndrome. Nat Med 10(12 Suppl):S88–S97. doi:10.1038/nm1143

4. Fehr AR, Perlman S (2015) Coronaviruses: an overview of their replication and pathogenesis. Methods Mol Biol 1282:1–23. doi:10.1007/978-1-4939-2438-7_1

5. Frieman M, Baric R (2008) Mechanisms of severe acute respiratory syndrome pathogenesis and innate immunomodulation. Microbiol Mol Biol Rev 72(4):672–685, Table of Contents. doi:10.1128/MMBR.00015-08

6. Liu DX, Fung TS, Chong KK, Shukla A, Hilgenfeld R (2014) Accessory proteins of SARS-CoV and other coronaviruses. Antiviral Res 109:97–109. doi:10.1016/j.antiviral.2014.06.013

7. Sulca T, Lindner HA, Purisima EO, Menard R (2005) Deubiquitination, a new function of the severe acute respiratory syndrome coronavirus papain-like protease? J Virol 79(7):4550–4551. doi:10.1128/JVI.79.7.4550-4551.2005

8. Oostra M, Hagemeijer MC, van Gent M, Bekker CP, te Lintelo EG, Rottier PJ, de Haan CA (2008) Topology and membrane anchoring of the coronavirus replication complex: not all hydrophobic domains of nsp3 and nsp6 are membrane spanning. J Virol 82(24):12392–12405. doi:10.1128/JVI.01219-08
9. Johnson MA, Chatterjee A, Neuman BW, Wutrich K (2010) SARS coronavirus unique domain: three-domain molecular architecture in solution and RNA binding. J Mol Biol 400(4):724–742. doi:10.1016/j.jmb.2010.05.027

10. Rack JG, Perina D, Ahel I (2016) Macrodomains: structure, function, evolution, and catalytic activities. Annu Rev Biochem. doi:10.1146/annurev-biochem-060815-014935

11. Teloni F, Altmeyer M (2016) Readers of poly(ADP-ribose): designed to be fit for purpose. Nucleic Acids Res 44(3):993–1006. doi:10.1093/nar/gkv1383

12. Feijs KL, Verheugd P, Luscher B (2013) Expanding functions of intracellular resident mono-ADP-ribosylation in cell physiology. FEBS J 280(15):3519–3529. doi:10.1111/febs.12315

13. Barkauskaite E, Jankevicius G, Ladurner AG, Ahel I, Timinszky G (2013) The recognition and removal of cellular poly(ADP-ribose) signals. FEBS J 280(15):3491–3507. doi:10.1111/febs.12358

14. Kusov Y, Tan J, Alvarez E, Enjuanes L, Hilgenfeld R (2015) A G-quadruplex-binding macrodomain within the “SARS-unique domain” is essential for the activity of the SARS-coronavirus replication-transcription complex. Virology 484:313–322. doi:10.1016/j.virology.2015.06.016

15. Paudyal S, Alfonso-Prieto M, Carnevale V, Redhu SK, Klein ML, Nicholson AW (2015) Combined computational and experimental analysis of a complex of ribonuclease III and the regulatory macrodomain protein, YmdB. Proteins 83(3):459–472. doi:10.1002/prot.24751

16. Forni D, Cagliani R, Mozzi A, Pozzoli U, Al-Daghri N, Clerici M, Sironi M (2016) Extensive positive selection drives the evolution of nonstructural proteins in lineage C Betacoronaviruses. J Virol 90(7):3627–3639. doi:10.1128/JVI.02988-15

17. Bacz-Santos YM, St John SE, Mesecar AD (2015) The SARS-coronavirus papain-like protease: structure, function and inhibition by designed antiviral compounds. Antivir Res 115:21–38. doi:10.1016/j.antiviral.2014.12.015

18. Fehr AR, Athmer J, Channappanavar R, Phillips JM, Meyerholz DK, Perlman S (2015) The nsP3 macrodomain promotes virulence in mice with coronavirus-induced encephalitis. J Virol 89(3):1523–1536. doi:10.1128/JVI.02596-14

19. Rose KM, Weiss SR (2009) Murine coronavirus cell type dependent interaction with the type I interferon response. Viruses 1(3):689–712. doi:10.3390/v1030689

20. Atkins GJ, Sheahan BJ (2016) Molecular determinants of alphavirus neuropathogenesis in mice. J Gen Virol. doi:10.1099/jgv.0.000467

21. Parvez MK (2015) The hepatitis E virus ORF1 ‘X-domain’ residues form a putative macrodomain protein/Appr-1”-phosphate catalytic-site, critical for viral RNA replication. Gene 566 (1):47–53. doi:10.1016/j.gene.2015.04.026

22. Koonin EV, Gorbalenya AE, Purdy MA, Rozanov MN, Reyes GR, Bradley DW (1992) Computer-assisted assignment of functional domains in the nonstructural polyprotein of hepatitis E virus: delineation of an additional group of positive-strand RNA plant and animal viruses. Proc Natl Acad Sci U S A 89(17):8259–8263

23. Saikatendu KS, Joseph JS, Subramanian V, Clayton T, Griffith M, Moy K, Velasquez J, Neuman BW, Buchmeier MJ, Stevens RC, Kuhn P (2005) Structural basis of severe acute respiratory syndrome coronavirus ADP-ribose-1”-phosphate dephosphorylation by a conserved domain of nsP3. Structure 13 (11):1665–1675. doi:10.1016/j.str.2005.07.022

24. Putics A, Filipowicz W, Hall J, Gorbalenya AE, Ziebuhr J (2005) ADP-ribose-1”-monophosphatase: a conserved coronavirus enzyme that is dispensable for viral replication in tissue culture. J Virol 79 (20):12721–12731. doi:10.1128/JVI.79.20.12721-12731.2005

25. Kuri T, Eriksson KK, Putics A, Zust R, Snijder EJ, Davidson AD, Siddell SG, Thiel V, Ziebuhr J, Weber F (2011) The ADP-ribose-1”-monophosphatase domains of severe acute respiratory syndrome coronavirus and human coronavirus 229E mediate resistance to antiviral interferon responses. J Gen Virol 92 (Pt 8):1899–1905. doi:10.1099/vir.0.031856-0

26. Hurst-Hess KR, Kuo L, Masters PS (2015) Dissection of amino-terminal functional domains of murine coronavirus nonstructural protein 3. J Virol 89(11):6033–6047. doi:10.1128/JVI.0197-15

27. Gorbalenya AE, Snijder EJ, Spaan WJ (2004) Severe acute respiratory syndrome coronavirus phylogeny: toward consensus. J Virol 78(15):7863–7866. doi:10.1128/JVI.78.15.7863-7866.2004

28. Jaroszewski L, Li Z, Cai XH, Weber C, Godzik A (2011) FFAS server: novel features and applications. Nucleic Acids Res 39(Web Server issue):W38–W44. doi:10.1093/nar/gkr441
29. Drozdetskiy A, Cole C, Procter J, Barton GJ (2015) JPred4: a protein secondary structure prediction server. Nucleic Acids Res 43(W1):W389–W394. doi:10.1093/nar/gkv332

30. Buchan DW, Minneci F, Nugent TC, Bryson K, Jones DT (2013) Scalable web services for the PSIPRED Protein Analysis Workbench. Nucleic Acids Res 41(Web Server issue):W549–W557. doi:10.1093/nar/gkt381

31. Jones DT, Cozzetto D (2015) DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics 31(6):857–863. doi:10.1093/bioinformatics/btu744

32. Wootten JC (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput Chem 18(3):269–285. doi:10.1016/0097-8485(94)85023-2

33. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB (2003) Protein disorder prediction. Structure 11(11):1453–1459. doi:10.1016/j.str.2003.10.002

34. Dosztanyi Z, Csizmok V, Tompa P, Simon I (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21(16):3433–3434. doi:10.1093/bioinformatics/bti541

35. Ishida T, Kinoshita K (2007) PrDOS: prediction of disordered protein regions from amino acid sequence. Nucleic Acids Res 35(Web Server issue):W460–W464. doi:10.1093/nar/gkm363

36. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y (2015) The I-TASSER suite: protein structure and function prediction. Nat Methods 12(1):7–8. doi:10.1038/nmeth.3213

37. Wang L, Brown SJ (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res 34(Web Server issue):W243–W248. doi:10.1093/nar/gkl298

38. Roy A, Yang J, Zhang Y (2012) COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. Nucleic Acids Res 40(Web Server issue):W471–W477. doi:10.1093/nar/gks372

39. Kallberg M, Wang H, Wang S, Peng J, Wang Z, Lu H, Xu J (2012) Template-based protein structure modeling using the RaptorX web server. Nat Protoc 7(8):1511–1522. doi:10.1038/nprot.2012.085

40. Wang S, Li W, Liu S, Xu J (2016) RaptorX-property: a web server for protein structure property prediction. Nucleic Acids Res. doi:10.1093/nar/gkw306

41. Engelhardt BE, Jordan MI, Srouji JR, Brenner SE (2011) Genome-scale phylogenetic function annotation of large and diverse protein families. Genome Res 21(11):1969–1980. doi:10.1101/gr.104687.109

42. Sahraian SM, Luo KR, Brenner SE (2015) SIFTER search: a web server for accurate phylogeny-based protein function prediction. Nucleic Acids Res 43(W1):W141–W147. doi:10.1093/nar/gkv461

43. Laskowski RA, Watson JD, Thornton JM (2005) ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res 33(Web Server issue):W89–W93. doi:10.1093/nar/gki414

44. Yang J, Roy A, Zhang Y (2013) Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. Bioinformatics 29(20):2588–2595. doi:10.1093/bioinformatics/btt447

45. Yang J, Roy A, Zhang Y (2013) BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. Nucleic Acids Res 41(Database issue):D1096–D1103. doi:10.1093/nar/gks966

46. Johnson M, Zaretzkaia I, Raytceis Y, Merezhuk Y, McGinnis S, Madden TL (2008) NCBI BLAST: a better web interface. Nucleic Acids Res 36(Web Server issue):W5–W9. doi:10.1093/nar/gkn201

47. Altschul S (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402. doi:10.1093/nar/25.17.3389

48. Berman HM (2000) The Protein Data Bank. Nucleic Acids Res 28(1):235–242. doi:10.1093/nar/28.1.235

49. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247(4):536–540. doi:10.1016/s0022-2836(05)80134-2

50. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punter M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44(D1):D279–D285. doi:10.1093/nar/gkv1344

51. Tatusov RL (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28(1):33–36. doi:10.1093/nar/28.1.33
52. Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A (2005) FFAS03: a server for profile—profile sequence alignments. Nucleic Acids Res 33(Web Server issue):W284–W288. doi:10.1093/nar/gki418

53. Cuff JA, Barton GJ (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins Struct Funct Genet 40(3):502–511. doi:10.1002/1097-0134(20000815)40:3<502::aid-prot170>3.0.co;2-q

54. Tan J, Vounhein C, Smart OS, Bricogne G, Bollati M, Kusov Y, Hansen G, Mesters JR, Schmidt CL, Hilgenfeld R (2009) The SARS-unique domain (SUD) of SARS coronavirus contains two macrodomains that bind G-quadruplexes. PLoS Pathog 5(5):e1000428. doi:10.1371/journal.ppat.1000428

55. Kong L (2004) Delineation of modular proteins: domain boundary prediction from sequence information. Brief Bioinform 5(2):179–192. doi:10.1093/bib/5.2.179

56. Bryson K, McGuffin LJ, Marsden RL, Ward JJ, Sodhi JS, Jones DT (2005) Protein structure prediction servers at University College London. Nucleic Acids Res 33(Web Server issue):W36–W38. doi:10.1093/nar/gki410

57. Wernisch L, Wodak SJ (2005) Identifying structural domains in proteins. 365–385. doi:10.1002/0471721204.ch18

58. Rack JG, Perina D, Ahel I (2016) Macrodomains: structure, function, evolution, and catalytic activities. Annu Rev Biochem 85:431–454. doi:10.1146/annurev-biochem-060815-104935

59. Egloff MP, Malet H, Putics A, Heinonen M, Dutartre H, Frangeul A, Gruez A, Campanacci V, Cambillau C, Ziebuhr J, Ahola T, Canard B (2006) Structural and functional basis for ADP-ribose and poly(ADP-ribose) binding by viral macro domains. J Virol 80(17):8493–8502. doi:10.1128/JVI.00713-06

60. Jankevicius G, Hassa PO, Lucas PO, Hottiger MO (2013) Macrodomain-containing proteins are new mono-ADP-ribosylhydrolases. Nat Struct Mol Biol 20(4):502–507. doi:10.1038/nsmb.2521

61. Karras GI, Kustatscher G, Buhecha HR, Allen MD, Pugieux C, Sait F, Bycroft M, Ladurner AG (2005) The macro domain is an ADP-ribose binding module. EMBO J 24(11):1911–1920. doi:10.1038/sj.emboj.7600664

62. Patel CN, Koh DW, Jacobson MK, Oliveira MA (2005) Identification of three critical acidic residues of poly(ADP-ribose) glycohydrolase involved in catalysis: determining the PARG catalytic domain. Biochem J 388(Pt 2):493–500. doi:10.1042/BJ20040942

63. Panda S, Poirier GG, Kay SA (2002) Tej defines a role for poly(ADP-Ribosyl)ation in establishing period length of the Arabidopsis circadian oscillator. Dev Cell 3(1):51–61. doi:10.1016/s1534-5807(02)00200-9

64. Varshavsky A (1997) The N-end rule pathway of protein degradation. Genes Cells 2(1):13–28. doi:10.1046/j.1365-2443.1997.1020301.x

65. Lu G, Liu D (2012) SARS-like virus in the Middle East: a truly bat-related coronavirus causing human diseases. Protein Cell 3(11):803–805. doi:10.1007/s13238-012-2811-1

66. Wassenaar TA, van Dijk M, Loureiro-Ferreira N, van der Schot G, de Vries SJ, Schmitz C, van der Zwan J, Boelens R, Giachetti A, Ferella L, Rosato A, Bertini I, Herrmann T, Jonker HRA, Baraia A, Jarvane V, Güntert P, Schwalbe H, Vranken WF, Doreleijers JF, Vriend G, Vister GH, Franke D, Kikhney A, Svergun DI, Fogh RH, Ionides J, Laue ED, Sprocket S, Jurkša S, Verlato M, Badoer S, Dal Pra S, Mazzucato M, Frizziero E, Bonvin AMJJ (2012) WeNMR: structural biology on the grid. J Grid Computing 10(4):743–767. doi:10.1007/s10723-012-9246-z

67. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF chimera—a visualization system for exploratory research and analysis. J Comput Chem 25(13):1605–1612. doi:10.1002/jcc.20084

68. Zhang Y, Skolnick J (2004) SPICKER: a clustering approach to identify near-native protein folds. J Comput Chem 25(6):865–871. doi:10.1002/jcc.20011

69. Brylinski M, Skolnick J (2008) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. Proc Natl Acad Sci U S A 105(1):129–134. doi:10.1073/pnas.0707684105
72. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput Biol 5(12):e1000585. doi:10.1371/journal.pcbi.1000585

73. de Beer TA, Berka K, Thornton JM, Laskowski RA (2014) PDBsum additions. Nucleic Acids Res 42(Database issue):D292–D296. doi:10.1093/nar/gkt940

74. Xu J, Li M, Kim D, Xu Y (2003) Raptor: optimal protein threading by linear programming. J Bioinforma Comput Biol 01(01):95–117. doi:10.1142/s0219720003000186

75. Peng J, Xu J (2011) A multiple-template approach to protein threading. Proteins 79(6):1930–1939. doi:10.1002/prot.23016

76. Cho CC, Lin MH, Chuang CY, Hsu CH (2016) Macro domain from Middle East respiratory syndrome coronavirus (MERS-CoV) is an efficient ADP-ribose binding module: CRYSTAL STRUCTURE AND BIOCHEMICAL STUDIES. J Biol Chem 291(10):4894–4902. doi:10.1074/jbc.M115.700542

77. Ahmad S, Sarai A (2005) PSSM-based prediction of DNA binding sites in proteins. BMC Bioinformatics 6:33. doi:10.1186/1471-2105-6-33

78. Neuman BW, Joseph JS, Saikatendu KS, Serrano P, Chatterjee A, Johnson MA, Liao L, Klaus JP, Yates JR 3rd, Wuthrich K, Klaus JP, Buchmeier MJ, Hsu CH (2016) Macro domain from Middle East respiratory syndrome coronavirus (MERS-CoV) is an efficient ADP-ribose binding module: CRYSTAL STRUCTURE AND BIOCHEMICAL STUDIES. J Biol Chem 291(10):4894–4902. doi:10.1074/jbc.M115.700542

79. Brylinski M, Skolnick J (2010) Comparison of structure-based and threading-based approaches to protein functional annotation. Proteins 79(6):1930–1939. doi:10.1002/prot.23016

80. Chen YC (2015) Beware of docking! Trends Pharmacol Sci 36(2):78–95. doi:10.1016/j.tips.2014.12.001

81. Pierce BG, Wiehe K, Hwang H, Kim BH, Vreven T, Weng Z (2014) ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. Bioinformatics 30(12):1771–1773. doi:10.1093/bioinformatics/btu097

82. Vreven T, Hwang H, Weng Z (2011) Integrating atom-based and residue-based scoring functions for protein-protein docking. Protein Sci 20(9):1576–1586. doi:10.1002/pro.687

83. Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/

84. van Zundert GC, Rodrigues JP, Trellet M, Schmitz C, Kastritis PL, Karaca E, Melquiond AS, van Dijk M, de Vries SJ, Bonvin AM (2016) The HADDOCK2.2 web server: user-friendly integrative modeling of biomolecular complexes. J Mol Biol 428(4):720–725. doi:10.1016/j.jmb.2015.09.014

85. Dominguez C, Boelens R, Bonvin AM (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc 125(7):1731–1737. doi:10.1021/ja026939x

86. Brunger AT, Adams PD, Clore GM, DeLano WL, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 54(Pt 5):905–921

87. Chatterjee A, Johnson MA, Serrano P, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 47(7):1739–1749. doi:10.1021/jm0306430

88. Friessner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS (2004) Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. J Med Chem 47(7):1750–1759. doi:10.1021/jm030644s

89. Chatterjee A, Johnson MA, Serrano P, Pedrini B, Joseph JS, Neuman BW, Saikatendu K, Buchmeier MJ, Kuhn P, Wuthrich K (2009) Nuclear magnetic resonance structure shows that the severe acute respiratory syndrome coronavirus-unique domain contains a macromolecular fold. J Virol 83(4):1823–1836. doi:10.1128/JVI.01781-08

90. Zhang Z, Dai J, Veliaith E, Jones RA, Yang D (2010) Structure of a two-G-tetrad intramolecular G-quadruplex formed by a variant human telomeric sequence in K+ solution: insights into the interconversion of human telomeric G-quadruplex structures. Nucleic Acids Res 38(5):1009–1021. doi:10.1093/nar/gkp1029

91. Hannig G, Makrides SC (1998) Strategies for optimizing heterologous protein expression in Escherichia coli. Trends Biotechnol 16(2):54–60. doi:10.1016/s0167-7799(97)01155-4
93. Skoog DA, Leary JJ (1992) Principles of instrumental analysis, 4th edn. Saunders College, Fort Worth
94. John Cavanagh WJF, Palmer AG III, Rance M, Skelton NJ (2007) Protein NMR spectroscopy: principles and practice. Academic Press, Amsterdam
95. Jacobsen NE (2007) NMR spectroscopy explained: simplified theory, applications and examples for organic chemistry and structural biology. Wiley, New York
96. Bieri M, Kwan AH, Mobli M, King GF, Mackay JP, Gooley PR (2011) Macromolecular NMR spectroscopy for the non-spectroscopist: beyond macromolecular solution structure determination. FEBS J 278(5):704–715. doi:10.1111/j.1742-4658.2011.08005.x
97. Billeter M, Wagner G, Wuthrich K (2008) Solution NMR structure determination of proteins revisited. J Biomol NMR 42(3):155–158. doi:10.1007/s10858-008-9277-8
98. Kay LE (2016) New views of functionally dynamic proteins by solution NMR spectroscopy. J Mol Biol 428(2 Pt A):323–331. doi:10.1016/j.jmb.2015.11.028
99. McDermott A (2009) Structure and dynamics of membrane proteins by magic angle spinning solid-state NMR. Annu Rev Biophys 38:385–403. doi:10.1146/annurev.biophys.050708.133719
100. Boehr DD, Dyson HJ, Wright PE (2006) An NMR perspective on enzyme dynamics. Chem Rev 106(8):3055–3079. doi:10.1021/cr050312q
101. Mayer M, Meyer B (1999) Characterization of ligand binding by saturation transfer difference NMR spectroscopy. Angew Chem Int Ed 38(12):1784–1788. doi:10.1002/ (s i c i)1521-3773(19990614)38:12<1784::aid-anie1784>3.0.co;2-q
102. Meyer B, Peters T (2003) NMR spectroscopy techniques for screening and identifying ligand binding to protein receptors. Angew Chem Int Ed Engl 42(8):864–890. doi:10.1002/anie.200390233
103. Konuma T, Lee YH, Goto Y, Sakurai K (2013) Principal component analysis of chemical shift perturbation data of a multiple-ligand-binding system for elucidation of respective binding mechanism. Proteins 81(1):107–118. doi:10.1002/prot.24166
104. Williamson MP (2013) Using chemical shift perturbation to characterise ligand binding. Prog Nucl Magn Reson Spectrosc 73:1–16. doi:10.1016/j.pnmrs.2013.02.001
105. Garner MM, Revzin A (1981) A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. Nucleic Acids Res 9(13):3047–3060. doi:10.1093/nar/9.13.3047
106. Greenfield NJ (2006) Using circular dichroism spectra to estimate protein secondary structure. Nat Protoc 1(6):2876–2890. doi:10.1038/nprot.2006.202
107. Srerama N, Woody RW (1993) A self-consistent method for the analysis of protein secondary structure from circular dichroism. Anal Biochem 209(1):32–44. doi:10.1006/abio.1993.1079
108. van Dijk M, Bonvin AM (2009) 3D-DART: a DNA structure modelling server. Nucleic Acids Res 37(Web Server issue):W235–W239. doi:10.1093/nar/gkp287