A Multilevel Framework for Analysis of Protein Folding Involving Disulphide Bond Formation

Patryk A. Wesołowski, David J. Wales, and Philipp Pracht*

Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

E-mail: pp555@cam.ac.uk

Abstract

Keywords: protein folding, BPTI, multiscale modelling, QM/MM, ONIOM, tight-binding

Abstract

In this study, a three-layered multi-centre ONIOM approach is implemented to characterise the naive folding pathway of bovine pancreatic trypsin inhibitor (BPTI). Each layer represents a distinct level of theory, where the initial layer encompassing the entire protein is modelled by a general all-atom force-field GFN-FF. An intermediate electronic structure layer consisting of three multicentre fragments is introduced with the state-of-the-art semiempirical tight-binding method GFN2-xTB. Higher accuracy, specifically addressing breaking and formation of the three disulphide bonds, is achieved at the innermost layer by using the composite DFT method r^2SCAN-3c. Our analysis sheds light on the structural stability of BPTI, particularly the significance of interlinking disulphide bonds. The accuracy and efficiency of the multicentre QM/SQM/MM approach are benchmarked using the oxidative formation of cystine. For the folding pathway of BPTI, relative stabilities are investigated through the calculation of free energy contributions for selected intermediates, focusing on the impact of the disulphide bond. Our results highlight the intricate trade-off between accuracy and computational cost, demonstrating that the multicentre ONIOM approach provides a well-balanced and comprehensive solution to describe electronic structure effects in biomolecular systems. We conclude that the multiscale energy landscape exploration provides a robust methodology for the study of intriguing biological targets.

1 Introduction

Proteins are macromolecular biopolymers essential to the structure and function of carbon-based organisms, governing most biochemical processes within cellular environments. These versatile biomolecules serve as catalysts, facilitate enzymatic reactions, and play important roles in cellular signalling, transport, and structural support. Their optimal functionality depends on a precisely folded native structure. In the past decade, significant scientific advancements, notably through endeavours such as The Critical Assessment of protein Structure Prediction (CASP) experiment,[1,2] showcased remarkable progress in predicting protein conformations. While the introduction of machine learning approaches, exemplified by the AlphaFold project,[3,4] marked a substantial leap forward, achieving an accurate description of thermodynamics and kinetics remains a key ambition in the realms of computational chemistry and biology.[5] Here, the folding process is a nuanced exploration of conformational space, aiming to identify the thermodynamically stable intermediates, and, ideally leading to the native state on the energy landscape.[6–10]

In this context, one of the most well-studied examples is the analysis of the folding pathway of bovine pancreatic trypsin inhibitor (BPTI) (Fig. 1). BPTI, a monomeric globular polypeptide, includes 16 distinct amino acids, with 58 residues folded into a stable and compact tertiary structure. The crystal structure reveals a twisted β-hairpin, C-terminal and N-terminal α-helices, and three disulphide bonds (Cys_{30}-Cys_{51}, Cys_5-Cys_{55}, Cys_{14}-Cys_{38}).[11] The stabilisation induced by disulphide bonds in the native state has been thoroughly explored in the literature.[12–15] The folding pathway energetically favours a structure with a well-defined tertiary arrangement. Intriguingly, the folding pathways encounter intermediate disulphide bonds (Cys_5-Cys_{30}, Cys_5-Cys_{14}, Cys_5-Cys_{38}, Cys_5-Cys_{51}), absent in the final native state.[14] BPTI's modest size and its well-established structural behaviour during folding render it an ideal target for scrutinising the thermodynamical stability of disulphide bonds and their impact on structural stability.

However, achieving high precision in calculations demands ingenuity, posing a significant

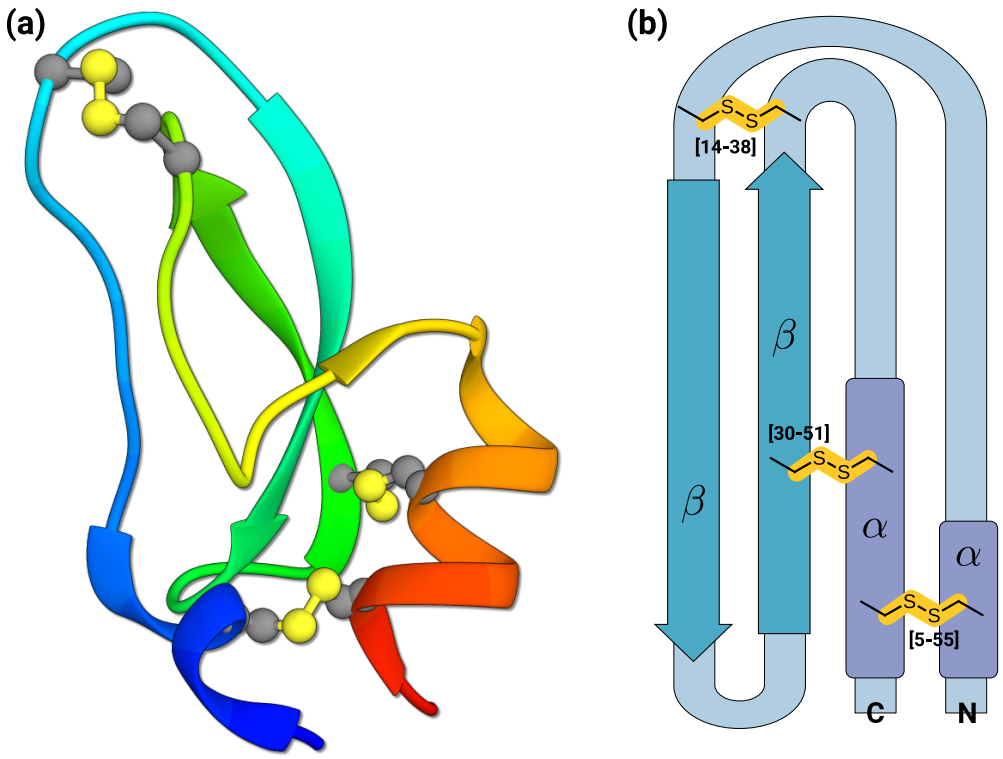

Figure 1: **(a)** Ribbon model of BPTI from the crystal structure with PDB ID 5PTI[11]. Cys-Cys sulphide bridges are shown in atomistic representation. **(b)** Simplified model of the BPTI protein, highlighting key secondary structure motifs and native disulphide bonds.

challenge in the modern computational modelling of proteins.[2,16–18] Unfortunately, there is a necessary trade-off between computational accuracy and efficiency. The scalability hurdle, where an increase in degrees of freedom leads to a disproportionate rise in computation time, renders quantum mechanical (QM) methods, such as density functional theory (DFT), mostly impractical for large biomolecular systems.[19,20] Much less costly classical force-field or coarse-grained methods[21] allow exploration of substantially larger systems, but often suffer in terms of accuracy and parameter availability. Balanced computational performance is essential, which in recent years has been addressed by machine learning (ML) methods in materials science[22,23] and the revival of semiempirical quantum mechanics (SQM).[24,25]

An alternative approach is the construction of multiscale models employing different levels of theory for different parts of the system, and, in particular, QM/MM schemes have emerged as viable for biomolecular systems.[26–29] Here, chemically important regions of a

system, for example, those exerting the greatest influence on structural stability or playing a pivotal role in occurring reactions, are selected as a "high layer" and assessed with a more accurate method (i.e. QM), while the remaining structure constitutes the "lower layer", calculated using molecular mechanics (MM) methods.

In this study, we implement a subtractive QM/MM scheme of the popular ONIOM (our own N-layered Integrated molecular Orbital and Molecular mechanics) type[30–34] to investigate relative structural stability in the BPTI folding pathway. Several studies on BPTI employing various QM/MM models can be found in the literature.[35–37] Focusing on the formation of the three disulphide bonds, we calculate the initial pathway to the native state with the general all-atom force-field GFN-FF.[38] A multiscale approach is then used to obtain free energies for selected points of interest along the folding path, employing DFT calculations at the r^{2}SCAN-3c level[39,40] to accurately describe the disulphide bond stability. As a central novelty, we introduce an intermediate SQM level between the DFT and GFN-FF layer to more accurately incorporate electronic structure effects in the vicinity of the cysteine residues. SQM methods of the tight-binding type[24] have successfully been employed in QM/MM simulations.[41] Tight-binding models of the GFNn-xTB family[25] have proved to be particularly versatile for exploration of energy landscapes,[42–44] and efficient calculations of thermodynamic properties.[20,45,46] and an ONIOM methodology has recently been adapted.[25,47] Our implementation of this multicentre ONIOM (MC-ONIOM) variant extends the approach, promising even greater computational time savings for large systems.[34]

We first present the multicentre and multilayer ONIOM approach, followed by a discussion on the simplest dipeptide containing a disulphide bond, cystine, to validate the theory behind MC-ONIOM for BPTI, followed by a discussion of the folding pathway. All the benchmarking calculations were performed in vacuum. This approach avoids additional complications associated with choice of a solvent model (explicit, implicit, or a hybrid of both) and associated sampling issues. Further, the choice of suitable implicit solvation models within the ONIOM context can be especially challenging and sometimes requires dedicated

implementations.[32] We plan to extend our tests to include solvation in future work. The constraints associated with the disulphide bond network enable us to focus on well-defined events on the folding pathway, which is a particular advantage for a benchmarking effort. We aim to gain insight on the role of disulphide bond stabilities in the BPTI folding pathways and highlight potential shortcomings of modelling these molecular rearrangements with classical force-field methods.

2 Theory and technical details

2.1 Multicentre *n*-layer ONIOM

We have implemented a multicentre n-layer ONIOM mechanical embedding method, closely following the approach presented by Seeber et al.[34] In this context, the ONIOM layer dependencies can be most effectively illustrated as a tree graph, exemplified in Fig. 2a. Each node corresponds to a substructure of the original system, and, with the exclusion of the initial structure, is explicitly linked to a parent node contingent upon its layer. The user is required to allocate atoms from the initial system to different nodes, with truncated bonds being automatically saturated through the linking atoms.

By statically placing the saturating link atom at $\mathbf{r}'_l$ along the vector of the cut bond $\mathbf{r}_a - \mathbf{r}_b$, we avoid introducing any additional degrees of freedom:

$$\mathbf{r}'_l = \mathbf{r}_b + k_{ab}(\mathbf{r}_a - \mathbf{r}_b) \ . \tag{1}$$

The appropriate value for the factor k_{ab} is determined by evaluating the ratio of the covalent radii[48] (R^{cov}) of the relevant atoms using

$$k_{ab} = \frac{R_b^{\mathrm{cov}} + R_H^{\mathrm{cov}}}{R_a^{\mathrm{cov}} + R_b^{\mathrm{cov}}} \ , \tag{2}$$

where the linking atoms are specified as hydrogen, and thus R_H^{cov} represents the covalent

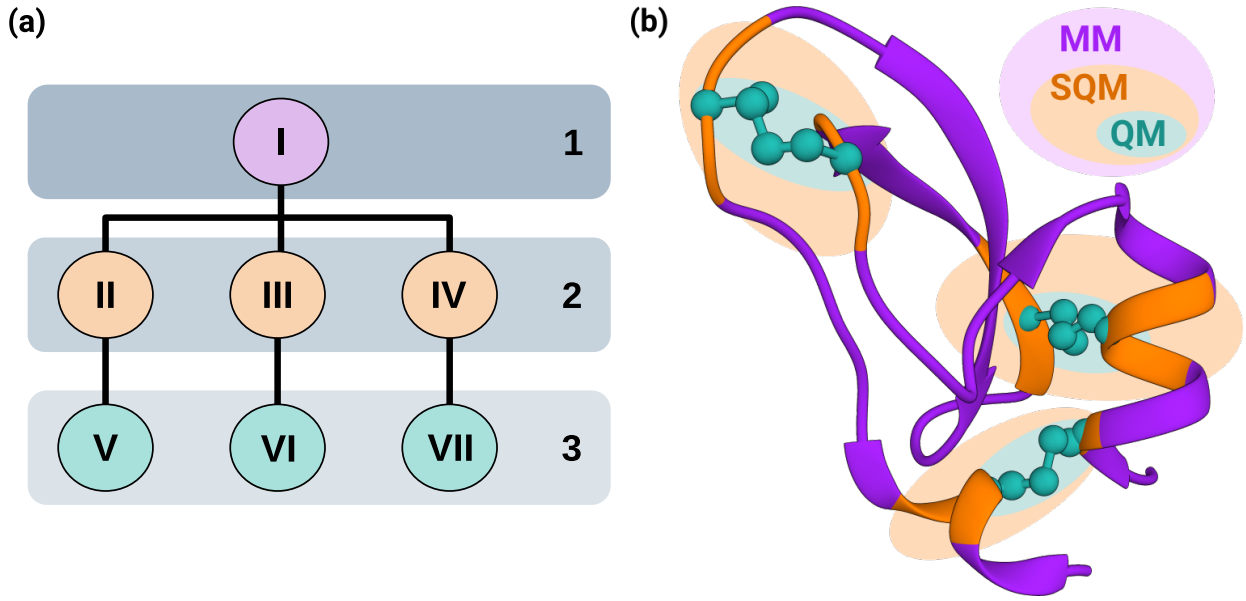

Figure 2: **(a)** Scheme detailing the construction of a three-layer MC-ONIOM dependency tree. Nodes in the diagram correspond to fragments, with the primary layer (Layer 1) requiring a solitary node encompassing all system atoms. Subsequent layers (Layers 2 and 3) have the flexibility to encompass various non-overlapping subsystems, functioning as child nodes of the initial node or nodes in subsequently higher layers. **(b)** Three-layer three-center ONIOM(r^2SCAN-3c:GFN2-xTB:GFN-FF) partitioning of the BPTI protein. Each ONIOM centre is built around one Cys-Cys sulphide bridge.

radius of hydrogen.

The final dependency tree facilitates the recursive assembly of the overall ONIOM properties. The general expression for building the energy, gradient, or Hessian, of a specific node, represented as $\mathcal{F}_i$, is given by:

$$\mathcal{F}_i = \mathcal{F}_i^h + \sum_j (\mathcal{F}_j - \mathcal{F}_j^l). \tag{3}$$

Here, $\mathcal{F}_i^h$ denotes the construction of the high-level energy/gradient for the parent node, F_j represents the recursively formed property of the child nodes, and $\mathcal{F}_j^l$ signifies their contribution to the low-level energy/gradient/Hessian. It is essential to note that the high-level calculations for each parent node align with the same level of theory as the low-level calculations for its respective child nodes. The recursion concludes when a node no longer has additional child nodes, at which point $\mathcal{F}_i = \mathcal{F}_i^h$. The gradient of the i-th model system $\mathbf{g}_i'$

can be projected into the basis of the real system $\mathbf{g}_i$ via

$$\mathbf{g}_i = \mathbf{g}'_i \mathbf{J}_i \ . \tag{4}$$

In a similar fashion, the Hessian matrix of the individual fragments $\mathbf{H}'_i$ can be projected into the basis of the real system via

$$\mathbf{H}_i = \mathbf{J}_i^\intercal \mathbf{H}'_i \mathbf{J}_i \ . \tag{5}$$

The Jacobian employed in both Eq. 4 and 5 is given by

$$\mathbf{J}_i = \begin{pmatrix} \mathbf{J}_{11} & \cdots & \mathbf{J}_{1n} \\ \vdots & \ddots & \vdots \\ \mathbf{J}_{m1} & \cdots & \mathbf{J}_{mn} \end{pmatrix}, \tag{6}$$

where m is the dimension of the i-th subsystem or fragment and n is the dimension of the real system. The corresponding matrix elements $\mathbf{J}_{nm}$ are given by

$$\mathbf{J}_{mn} = \mathbf{E} \cdot \frac{\partial \mathbf{r}'_m}{\partial \mathbf{r}_n}, \tag{7}$$

where $\mathbf{E}$ is a 3×3 identity matrix. The derivatives $\frac{\partial \mathbf{r}'_m}{\partial \mathbf{r}_n}$ are either 1, 0, or a value depending on k_{ab} due to Eq. 1. Once all gradients (or Hessians) of the subsystems have been projected into the basis of the real system, constructing the full MC-ONIOMn gradient (or Hessian) is possible via the recursive algorithm given by Eq. 3.

In this project, we employ a three-layers MC-ONIOM to investigate structures sampled from the BPTI folding pathway. The corresponding multiple centres are illustrated in Fig 2b, where each layer corresponds to a distinct level of accuracy. The initial layer provides the assessment of the entire structure at a classical force-field level, a task accomplished using GFN-FF.[38] The second, intermediate layer comprises three multicentre fragments, each addressed with GFN2-xTB.[49] The third and final layer includes only the three Cys-Cys

sulphide bridges and is computed using r^{2}SCAN-3c.[39,40]

2.2 Technical details

The MC-ONIOM approach was implemented in a standalone Fortran library called lwONIOM, which is freely available under the LGPL-3.0 license from GitHub.[50,51] The library provides bookkeeping and ONIOM partitioning functionalities (*cf.* Sec. 2.1) for mechanical embedding but, adhering to the eponymous "light-weight" in the coding of lwONIOM, no potential calculators (energies and gradients) are implemented at the time of writing. To provide this capability, lwONIOM was interfaced in the recently developed CREST program.[42,51,52] r^{2}SCAN-3c DFT calculations were performed with the ORCA program package.[53,54] SQM and MM calculations were conducted with dedicated libraries implemented in the CREST program, employing GFN2-xTB[25,49] and GFN-FF[25,38] methodologies, respectively. If not stated otherwise, we will refer to r^{2}SCAN-3c[39,40] as "DFT", to GFN2-xTB[49] as "SQM", and to GFN-FF[38] as "FF". In particular, two-layer ONIOM(DFT:SQM), and three-layer ONIOM(DFT:SQM:FF), are assessed with regard to their performance and computation times. Following standard conventions,[32,34] these approaches are denoted as (MC-)ONIOMn, with n referring to the number of layers. These levels of theory were first evaluated for the formation of cystine, where we compared reaction energies (ΔE) and free energies at 298.15 K ($\Delta G^{298.15}$). For reference, high-level DFT calculations were conducted with the range-separated ωB97X-V functional[55] employing a polarised quadruple-ζ basis set def2-QZVPP.[56] All DFT calculations in this study were conducted using the *TightSCF* and *DefGrid3* settings in ORCA. Calculation of the free energy employed the modified rigid-rotor-harmonic-oscillator (mRRHO) approximation for $\delta G_{\mathrm{mRRHO}}$ contributions, using Grimme's rovibrational entropy interpolation for frequencies less than 25 cm^{-1}.[57,58] The corresponding calculations were performed with the CREST code.

For BPTI, CREST was first employed to prepare the unfolded linear state and optimize both unfolded and native structures. The OPTIM program,[59] interfaced with GFN-FF, was

used to facilitate the calculation of the pathway between the unfolded and native states. Importantly, all pathway calculations for BPTI associated with the GFN-FF level of theory employed the same reference topology of the native state to enable the correct cleavage and (re-)formation of disulphide bonds, since the *a priori* creation of new bonds is not possible in the current formulation. The geometry optimisation techniques employed have been reviewed before.[6,7,60] Very briefly, transition state (TS) candidates are obtained using the doubly-nudged[61,62] elastic band[63–66] (DNEB) method. We first obtained an initial database of stationary points by optimising all structures from an approximate pathway obtained via quasi-continuous interpolation (QCI).[67,68] The collected minima were then connected pairwise via DNEB, employing up to 150 images each, and TS candidates and new minima were added back to the database from which a complete pathway was identified. From this pathway, we selected the TS and associated minima describing the three disulphide bond cleavages. These minima were subjected to a detailed analysis of their relative stabilities using the MC-ONIOM approach to calculate free energy corrections from the reconstructed (Eq. 3 and 5) Hessian within the mRRHO approximation, as described above. For simplicity, all calculations were **originally** performed in the gas-phase, as explained in the Introduction. **Again, since one goal is to highlight the computational methodology, referring to vacuum calculations removes one layer of complexity (choice of solvent model, parametrisation, etc.) and allows us to focus on the calculation strategy. However, where feasible the gasphase results are supplemented by calculations performed with GFN-FF and the ALPB implicit solvation model[69] for water.**

3 Results and discussion

3.1 Cystine

Initially, we benchmarked our methodology on a simple biomolecular system containing a disulphide bond, specifically the cystine dipeptide. The dimeric structure is formally created by the oxidative reaction of two cysteine monomers.[70] Calculations of ΔE and $\Delta G^{298.15}$ were conducted following the reaction outlined in Fig. 3a, employing various levels of accuracy: GFN2-xTB, GFN-FF, ONIOM(DFT:SQM), and ONIOM(DFT:SQM:FF), as well as r^{2}SCAN-3c itself, and a high-accuracy ωB97X-V/def2-QZVPP//r^{2}SCAN-3c level.

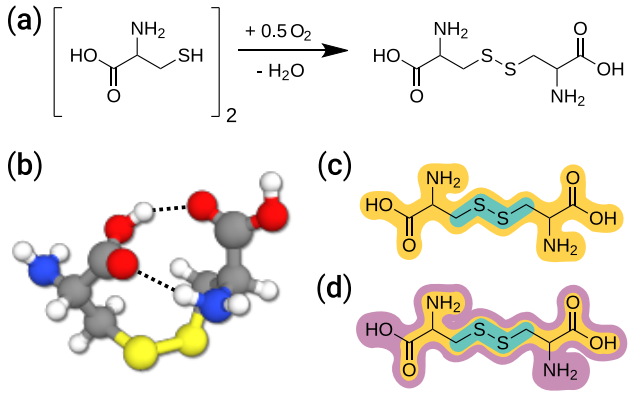

Figure 3: **(a)** Formation reaction of cystine from two cysteine monomers. **(b)** Most favourable gas phase conformer of cystine, calculated at the GFN-FF level. **(c)** Schematic two-layer ONIOM(r^{2}SCAN-3c:GFN2-xTB) setup. **(d)** Schematic three-layer ONIOM(r^{2}SCAN-3c:GFN2-xTB:GFN-FF) setup.

Table 1 presents a thorough examination of gas-phase reaction energies (ΔE), free energies at 298.15K ($\Delta G^{298.15}$), and computational efficiency across diverse methods applied to assess the oxidation of two cysteine monomers to cystine (*cf.* Fig. 3a). The reference method, ωB97X-V/def2-QZVPP//r^{2}SCAN-3c, establishes a baseline with a ΔE of $-75.00\,\mathrm{kcal\ mol^{-1}}$and $\Delta G^{298.15}$ of $-62.87\,\mathrm{kcal\ mol^{-1}}$. The range-separated ωB97X-V functional[55], known for its performance in established benchmark databases like GMTKN55,[71,72] serves as an excellent reference and is crucial for evaluating alternative methods. Unfortunately, its high computational cost renders it unsuitable for use in large-scale ONIOM

Table 1: Gas phase reaction energies (ΔE), free energies at 298.15 K ($\Delta G^{298.15}$), and cumulative CPU times per singlepoint calculation for the oxidation of two cysteine monomers to cystine (*cf.* Fig. 3a), calculated at several levels of theory. DFT refers to r^{2}SCAN-3c, SQM refers to GFN2-xTB, and FF refers to GFN-FF.

Method	**CPU Time** [s]	ΔE [**kcal mol^{-1}**]	$\Delta G^{298.15}$ [**kcal mol^{-1}**]
GFN-FF	0.03	−17.51	−2.21
GFN2-xTB	0.09	−95.28	−81.98
ONIOM3(DFT:SQM:FF)	19.83	−91.81	−78.03
ONIOM2(DFT:FF)	**19.92**	**−101.23**	**−86.60**
ONIOM2(DFT:SQM)	20.10	−85.69	−72.49
r^{2}SCAN-3c	49.08	−78.31	−65.97
ωB97X-V/def2-QZVPP//DFT	731.88	−75.00	−62.87

setups requiring hundreds of energy and gradient evaluations. r^{2}SCAN-3c, when compared to the ωB97X-V reference, exhibits a closely aligned ΔE of −78.31 kcal mol^{-1}, $\Delta G^{298.15}$ of −65.97 kcal mol^{-1}, and a balanced computation time, making it a viable choice within the ONIOM scheme. Furthermore, r^{2}SCAN-3c serves as the level for geometry optimization of the reference method. Frequencies calculated with ωB97X-V/def2-QZVPP on these structures were checked for the absence of imaginary modes, confirming this to be an adequate choice.

A noteworthy observation is the computational efficiency of the GFN-FF method, with a CPU time of 0.03 seconds. However, this efficiency comes with a trade-off, with differences in ΔE (−17.51 kcal mol^{-1}) and $\Delta G^{298.15}$ (−2.21 kcal mol^{-1}) from the reference. These observations are expected for a classical force-field method. Nonetheless, with a $\delta G_{\mathrm{mRRHO}}$ of approximately 15.30 kcal mol^{-1}, GFN-FF demonstrates its suitability[46] to calculate free energy contributions, which are on the same order of magnitude as for the higher-level reference methods. In contrast, the semiempirical electronic structure method GFN2-xTB achieves a substantially better ΔE of −95.28 kcal mol^{-1} and $\Delta G^{298.15}$ of −81.98 kcal mol^{-1}, albeit for a longer CPU time of 0.09 seconds per single-point calculation. The differences in the computational cost of GFN2-xTB and GFN-FF are insignificant for the cystine model system, but can become an important factor for larger biomolecules like BPTI.

The ONIOM3(DFT:SQM:FF) and ONIOM2(DFT:SQM) methodologies exhibit intermediate results for ΔE and $\Delta G^{298.15}$ when compared with GFN2-xTB and the DFT references. Specifically, ONIOM3 yields values of $-91.81\,\mathrm{kcal\ mol^{-1}}$ and$-78.03\,\mathrm{kcal\ mol^{-1}}$ for ΔE and $\Delta G^{298.15}$, while for ONIOM2 these values were $-85.69\,\mathrm{kcal\ mol^{-1}}$ and $-72.49\,\mathrm{kcal\ mol^{-1}}$, respectively. **Omission of the SQM "middle" layer, i.e., using an ONIOM2(DFT:FF) approach, does not improve on ONIOM2(DFT:SQM) nor ONIOM3 and strongly overestimates the cystine stability with an ΔE of $-101.23\,\mathrm{kcal\ mol^{-1}}$, and $\Delta G^{298.15}$ of $-86.60\,\mathrm{kcal\ mol^{-1}}$.** Additionally, **all three calculations show similar computational cost with around 20 seconds** of CPU time per singlepoint calculation**, which shows that t**hese methods strike a balance between accuracy and computational cost **but ultimately correlate with the level of theory.** The seemingly superior accuracy of ONIOM2 compared to ONIOM3 can be attributed to two factors: ONIOM3 incorporates less accurate GFN-FF results, impacting overall precision, and the choice of ONIOM subsystems proves sub-optimal for ONIOM3, especially regarding different setups for the two cysteine units. **Essentially, the ONIOM3 setup omits important regions of non-covalent interactions and thus introduces additional errors in the cystine test case.** This is a well-known challenge for QM/MM schemes, particularly for calculations within the supermolecular approach.[32,33] In the context of BPTI, this result demonstrates that regions close to the disulphide bonds *must* be described at the same level to avoid truncating non-covalent interactions. Therefore, in the following section, the cysteine units will be described at the DFT level, while all adjacent residues are modelled with GFN2-xTB. The remaining residues are modelled by the low-level force-field.

In summary, while GFN-FF excels in computational efficiency, ONIOM schemes strike a useful balance between accuracy and efficiency. The discrepancy between ONIOM2 and ONIOM3 underscores the significance of methodological considerations, including the choice of subsystems, in achieving accurate results. Compared to the reference method ωB97X-V/def2-QZVPP, r^{2}SCAN-3c proves to be a suitable choice for the ONIOM high-level calcu-

lations, providing reliable and cost-efficient ΔE and $\Delta G^{298.15}$ values.

3.2 BPTI

Before studying the disulphide bond stability via an ONIOM approach, we employed GFN-FF in conjunction with the OPTIM program to elucidate the folding pathway of BPTI. GFN-FF showcases robustness over a broad variety of chemical systems,[38] making it well-suited for applications involving elements up to radon, $Z \leq 86$. A limitation arises in the inability of GFN-FF to represent the formation of new bonds. Nevertheless, GFN-FF, functioning as a dissociative force-field, can model the homolytic rupture and reformation of disulphide bonds, which was exploited for the pathway construction using the DNEB approach (§2.2). This approach necessitates a formal force-field setup from the unfolding of the native state employing the native topology, i.e., containing the disulphide bonds, in all the calculations. The corresponding pathway is shown in Fig. 4. The integrated path length is approximately 250 Å, with the most pronounced energetic stabilisation occurring in the final third of the pathway, coinciding with the formation of the disulphide bonds. **We note at this point that local minima along the pathway may differ with and without consideration of solvation effects. However, since the DNEB approach starts from two defined endpoints (i.e., the unfolded and native state of BPTI) and the initial interpolation does not know about the "true" potential, we can reasonably expect that the overall folding trajectory will look similar. Nonetheless, large effects onto the energetics must be expected upon inclusion of a solvent. A singlepoint energy re-evaluation of the entire pathway at the GFN-FF/ALPB(H_2O) level of theory therefore reveals typical effects of the implicit salvation: Generally, implicit solvation attenuates non-covalent interactions,[45] which in this case manifests in an overall lower ΔE between folded and unfolded conformations, i.e., the relative energies in implicit solvation are lower compared to the gas phase energies in Fig. 4. Furthermore, unfolded conformation are more strongly affected by the**

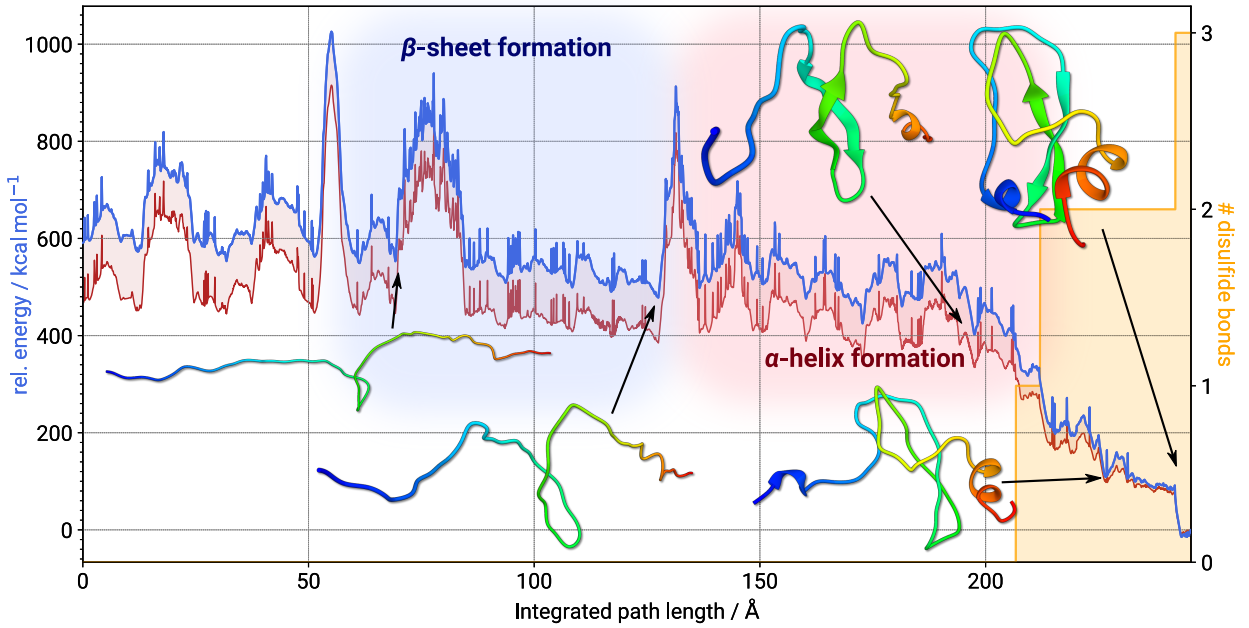

Figure 4: Folding pathway **(in blue)** for the BPTI protein showing the energy scale on the left, and number of disulphide bonds **(in yellow)**on the right. **The second energy curve plotted in red corresponds to singlpoint energy evaluations of the gasphase folding pathway re-evaluated with ALPB(H_2O) implicit solvation.** Sections of the pathway where the formation of β-sheet and α-helix motif occurs, are shaded in blue and red, respectively.

implicit solvation due to their larger surface area (to which part of the implicit solvation energy is proportional[69]). Secondary structure folding events commence with the formation of the β-sheet, succeeded by the development of the C-terminal helix and N-terminal helix, respectively. This hierarchy is reflected by the disconnectivity graph[73,74] for the stationary point database corresponding to the pathway, which is shown in Fig. 5a. Here, distinct funnels for the secondary motives and the main funnel containing the disulphide bond formation are clear. The disconnectivity graph in Fig. 5a is coloured to highlight the pathways leading to the creation of subsequent disulphide bonds. These sections of the pathway are associated with the three transition states TS1, TS2, and TS3, in order of unfolding, where the Cys_5-Cys_{55} disulphide bond makes or breaks in TS1, Cys_{30}-Cys_{51} in TS2, and Cys_{14}-Cys_{38} in TS3. Magnified sections of the corresponding pathways of TS1, TS2, and TS3 are shown in Fig. 5b, with the energy origin given relative to the starting minimum of

the TS1 pathway.

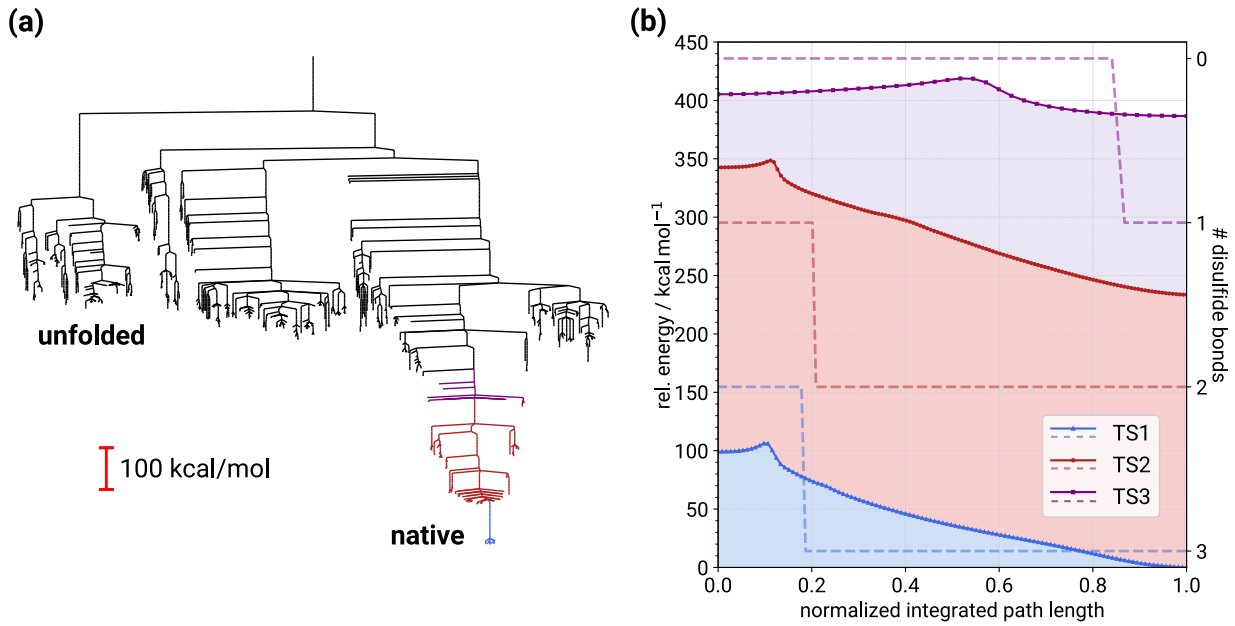

Figure 5: **(a)** Disconnectivity graph at the GFN-FF level, corresponding to the path in Fig. 4. **(b)** Pathways for the homolytic cleavage of the Cys_5-Cys_{55} (TS1), Cys_{30}-Cys_{51} (TS2), Cys_{14}-Cys_{38} (TS3) disulphide bonds, calculated at the GFN-FF level. Energies are shown relative to the starting point minimum of the TS1 pathway.

Due to the technical setup of the DNEB method, our pathway deviates from the sequence of disulfide cleavage events outlined in the literature.[14,75] Darby et al. proposed a pathway where the Cys_{30}-Cys_{51} bond forms first, followed by some alternative Cys-Cys combinations, Cys_5-Cys_{55}, and finally Cys_{14}-Cys_{38}. On the other hand, while our calculations agree in that Cys_{30}-Cys_{51} (TS2) forms prior to Cys_5-Cys_{55} (TS1), Cys_{14}-Cys_{38} (TS3) forms first in our calculation while it forms last according to the (most probable) experimentally observed pathway. The latter is an artifact of the DNEB interpolation which first "curls" the backchain of the protein, before bringing the N and C terminus closer together (*cf.* Fig. 4). These results demonstrate the technical capabilities of the DNEB approach in combination with GFN-FF. For the purpose of computational efficiency we focus on this single pathway to study the relative disulphide bond stability in the following. If the pathway is broken into smaller sections and alternative endpoints

are explored, this approach can be adapted for reproduction of competing disulphide bond formation from the experiment. Unfortunately, the force-field setup prevents us from exploring the alternative pathways that lead to the formation of non-native disulfide bonds intermediates, such as Cys_5-Cys_{14} and Cys_5-Cys_{38}. Since the corresponding disulphide bonds are not present in the topology, any force at between the sulfur atoms at close distance would be strongly repulsive. It is of course possible to model the individual pathways via standalone DNEB calculations, but connecting them on a single energy surface will require some code modifications to the force-field to allow manual definition of possible bonds. Nonetheless, previous studies indicate that the rearrangement process is unaffected by non-native disulfide bonds, as they tend to occur in relatively unfolded regions of the molecule.[75] The corresponding bonds are not deemed "committed" to the rearrangement process, suggesting their occurrence post the rate-determining step.[76] Hence, we believe that the naive folding pathway is still an adequate representation to showcase the DNEB and ONIOM methodologies and for judging the relative disulphide bond stability via a supramolecular approach.

Concerning the overall disulphide bond stability, the primary question is the physical nature of the GFN-FF pathway. **In fact, some problems may result due to the construction of the GFN-FF potential: The actual cleavage of bonds has no barrier in GFN-FF since the binding potential is modeled with a Gaussian function.[38] The transition states shown in Fig. 5b therefore correspond to the first conformational rearrangement after the disulphide bond cleavage, rather than the bond breaking itself.** In essence, the pathway lacks physical validity due to the aforementioned homolytically broken disulphide bonds. Formally, the cleaved cysteine residues are treated as radicals, constituting a chemical misrepresentation. In an aqueous environment, cysteine residues, contrary to this assumed behaviour, exhibit disulphide formation

more closely aligned with the oxidative reaction associated with the cystine formation, as modelled in §3.1. In a biological context, the corresponding reaction is driven by disulphide isomerase.[77] Consequently, the cleaved disulphide bonds necessitate the saturation of sulphur atoms with hydrogen to form thiols. To faithfully represent this scenario, reaction free energies ΔG, as employed previously for cystine, are utilised and further elucidated in a supramolecular approach. **The supramolecular approach herein enables us to focus on differences in relative stability rather than pathways that anyways might lack physical validity.**

Within the broader context of our study, we focus on the significance of the disulphide bond formation through modelling homolytic cleavage at GFN-FF level. Our findings suggest a useful model, featuring three distinct transition states associated with significant barriers along the overall pathway. Qualitatively, the disconnectivity graph of Fig. 5a suggests an energy difference between the native (NS) and homolytically unfolded (UF) states of around 600 kcal mol^{-1}. The disulphide bond funnel alone contributes approximately 400 kcal mol^{-1}, constituting roughly two-thirds of this energy difference. Clearly, the formation of these bonds is expected to contribute significantly to the stability of folded BPTI. Table 2 presents the reaction energies (ΔE) and free energies at 298.15 K ($\Delta G^{298.15}$) for three disulphide bond cleavages, calculated using MC-ONIOM3(DFT:SQM:FF) and GFN-FF. In evaluat-

Table 2: Reaction energies (ΔE) and free energies at 298.15 K ($\Delta G^{298.15}$) for the three disulphide bond cleavages calculated with MC-ONIOM3(DFT:SQM:FF) and GFN-FF. The ONIOM calculation and the GFN-FF(sat.) reaction refer to formal oxidation, as for cystine in Fig. 3a, GFN-FF(unsat.) refers to the pathways from Fig. 5b with homolytic disulphide bond cleavage.

States	**MC-ONIOM3**		**GFN-FF (sat.)**		**GFN-FF (unsat.)**	
	ΔE [**kcal mol^{-1}**]	$\Delta G^{298.15}$ [**kcal mol^{-1}**]	ΔE [**kcal mol^{-1}**]	$\Delta G^{298.15}$ [**kcal mol^{-1}**]	ΔE [**kcal mol^{-1}**]	$\Delta G^{298.15}$ [**kcal mol^{-1}**]
TS1	55.64	63.57	−10.88	−5.79	91.64	90.26
TS2	34.30	30.93	−28.99	−25.95	99.90	80.64
TS3	36.45	47.85	−56.52	−41.46	17.76	17.79
$\sum$	126.38	142.35	−96.40	−73.20	209.30	188.69
NS / UF	628.92	410.46	356.19	314.56	565.98	492.88

ing the reactions within the MC-ONIOM3 framework, both the changes in energy (ΔE) and the corresponding $\Delta G^{298.15}$ align with expected values for disulphide bond cleavages. As evidenced by comparisons with cystine (*cf.* section 3.1), both reaction energies and free energy contributions are of the correct order of magnitude, attributing to their physical plausibility. The three transition states corresponding to disulphide bond formation/cleavage, TS1, TS2, and TS3, exhibit ΔE values of 55.64, 34.30, and 36.45 kcal mol^{-1}, respectively, while the corresponding $\Delta G^{298.15}$ values are 63.57, 30.93, and 47.85 kcal mol^{-1}. The stability of the first disulphide bond Cys_5-Cys_{55} is notably greater than the subsequent values for TS2 and TS3. One rationale for this observation can be attributed to the influence of non-covalent interactions. TS1 is positioned "early" in **our simulation of** the cleavage pathway, exhibiting a predominantly folded configuration that requires disruption of numerous non-covalent before bond rupture. Further rotational barriers, particularly linked to α and β motives, exert a stabilizing influence through intramolecular non-covalent interactions. As a result, although the representation of disulphide bonds may account for part of the energy difference, the overall structural stability results from a combination of multiple contributing factors. **Obviously, a free energy difference can not be compared with observable rates. However, through experimental measurements it was confirmed that the quasi-native (Cys_5-Cys_{55}, Cys_{14}-Cys_{38}) intermediate is very stable whereas the Cys_{30}-Cys_{51} conformation requires recombination and formation of non-native disulphide bonds first.[75] We see this observation qualitatively confirmed by the relative stabilities calculated via the supramolecular approach and MC-ONIOM3: TS1 (Cys_5-Cys_{55}) has the largest $\Delta G^{298.15}$, followed by TS3 (Cys_{14}-Cys_{38}), while TS2 (Cys_{30}-Cys_{51}) is least stable. Furthermore, TS3 was observed to form rapidly,[14,75] which is in line with it having the smallest activation barrier among the three cleavages (*cf.* Fig. 5b).**

A discrepancy arises for GFN-FF following the same (chemically correct) supramolecular approach, as the sign of the reaction energies is incorrect. This inconsistency is expected

and underlines the intrinsic limitations of force-field methods in capturing the intricacies of chemical reactions. However, an interesting observation emerges when disulphide bond cleavage is approached homolytically, surprisingly revealing a correct sign. Here, the stabilities of TS1 and TS2 are overestimated compared to the underestimated TS3, indicating an overstabilization due to non-covalent interactions within the force-field. Nonetheless, energies and free energies are on a similar scale to MC-ONIOM3 and differ only by a factor of two to three. The cumulative values for the three reactions are also provided, emphasising the substantial differences in energy and free energy between the methodologies and defining the overall importance of the disulphide bonds. For MC-ONIOM3, the cumulative energy contribution contributes to roughly 20.1 % of the total NS/UF difference, and approximately one-third (33.9 %) of the free energy difference. The homolytically cleaving GFN-FF calculations estimate the same contributions as 37.0 % and 38.3 %, respectively. These results clearly demonstrate the relevance of the disulphide bonds for the BPTI folding process, although GFN-FF, again, qualitatively overestimates the corresponding contribution. Overall, the results presented in Table 2 underscore the importance of the chosen computational approach in determining the correct energetics of disulphide bond cleavages and emphasize the shortcomings of classical force-fields. Deliberately modelling a chemically implausible pathway, i.e. with homolytic disulphide bond ruptures, may provide a useful computational strategy to model the folding process. However, we strongly recommend investigating this possibility on a case-by-case basis. **Furthermore, the ONIOM framework provides an exceptional strategy to obtain accurate energetics at lower-than-DFT cost. In particular, the Hessian recombination according to Eq. 5 enables great computation time savings for ΔG calculations. Here, the seminumerical calculation at GFN-FF level took 10 min 11 sec, while the corresponding MC-ONIOM3 calculation took 146 min 32 sec (both on 10 CPUs of the same machine). The latter is identical to the accumulation of all ONIOM subsystem Hessian calculations and therefore is much cheaper than even a singlepoint energy of the entire system**

at DFT level, which we were unable to obtain within a 24h job time limit.

As noted above, the inclusion of solvent effects can strongly influence the calculated energetics. Therefore, reaction (free) energies were also calculated employing GFN-FF with the ALPB(H_2O) implicit solvation model and are given in Table 3. For the respective MC-ONIOM3 calculations, only the GFN-FF layer experiences the implicit solvation potential. The efficacy of this approach will be investigated in future studies, however, as an approximate treatment it is sufficient to at least qualitatively capture some of the solvation effects. Generally, the energy difference between folded and unfolded state of BPTI is reduced due to the implicit solvation. As noted earlier, this is due to an attenuation of non-covalent interactions and a lowering of the unfolded state energies relative to the native state, presumably due to the large solvent accessible surface area. For MC-ONIOM3, energy differences of TS1 and TS2 are slightly greater, and for TS3 slightly lower than the corresponding gasphase calculations. Herein, free energy contributions for TS1 and TS3 further stabilize the corresponding disuphide bonds (i.e., $\Delta G^{298.15}$ is greater than the respective ΔE), while an opposite trend is observable for TS2. This is again qualitatively in line with experimen-

Table 3: Reaction energies (ΔE) and free energies at 298.15 K ($\Delta G^{298.15}$) for the three disulphide bond cleavages calculated with MC-ONIOM3(DFT:SQM:FF/ALPB(H_2O)) and GFN-FF/ALPB(H_2O). The ONIOM calculation and the GFN-FF(sat.) reaction refer to formal oxidation, as for cystine in Fig. 3a, GFN-FF(unsat.) refers to the pathways from Fig. 5b with homolytic disulphide bond cleavage. Within the ONIOM calculation ALPB implicit solvation was employed only for the outer force-field layer.

States	MC-ONIOM3		GFN-FF/ALPB (sat.)		GFN-FF/ALPB (unsat.)	
	ΔE [kcal mol^{-1}]	$\Delta G^{298.15}$ [kcal mol^{-1}]	ΔE [kcal mol^{-1}]	$\Delta G^{298.15}$ [kcal mol^{-1}]	ΔE [kcal mol^{-1}]	$\Delta G^{298.15}$ [kcal mol^{-1}]
TS1	75.97	79.95	1.92	5.52	39.09	40.18
TS2	59.20	33.03	-15.07	-0.54	43.37	37.26
TS3	34.81	57.00	-37.50	-20.89	25.29	33.93
$\sum$	169.97	169.98	-50.65	-15.91	107.75	111.36
NS / UF	429.42	371.61	281.58	235.96	413.22	350.56

tal[75] observations: The large $\Delta G^{298.15}$ of TS1 and TS3 can explain the formation of a very stable quasi-native (Cys_5-Cys_{55}, Cys_{14}-Cys_{38}) structure, while the Cys_{30}-Cys_{51} is a partly-folded intermediate and thus should show a smaller free energy difference for the disulphide bond formation. On the other hand, calculations performed at the GFN-FF/ALPB(H_2O) (sat.) level once more exhibit incorrect signs for TS2 and TS3, further highlighting the limitations of the force-field method when applied in supramolecular approaches. Similarly, the artificial homolytic cleavage structures calculated at the same level of theory again provide estimates surprisingly close to the reference MC-ONIOM3 approach. The contribution of disulphide bond stability to the NS/UF difference is significantly larger ($\approx 45\%$ for $\Delta G^{298.15}$ at the MC-ONIOM3 level) than in the gas phase. This underscores the critical role of covalent disulphide bonds relative to the dampened non-covalent interactions of the protein in solution. Solvent effects play a crucial role in assessing this stability. However, considering that these effects are most pronounced in the unfolded regions of the pathway (*cf.* Fig. 4), rather than in the regions where disulphide bonds form, we will revert to vacuum calculations in subsequent discussions to bypass further complexities.

Further considerations must be made concerning the structural stability and conformational changes in the studied system due to the level of theory choice. Hence, Fig. 6 provides a visual comparison between the MC-ONIOM3 and GFN-FF reoptimised structures for the minima associated with the paths TS1, TS2, and TS3. The heavy-atom root-mean-square deviations (RMSD) are displayed below each corresponding structure pair. This comparison allows for an assessment of the structural deviations between the two computational approaches. RMSDs underscore the crucial role of both non-covalent interactions and disulphide bonds in determining structural outcomes. The structure is well defined by the disulphide bonds, as indicated by the remarkably low heavy-atom RMSD for the two minima connected by TS1. Energies and free energies calculated for these structures will be consis-

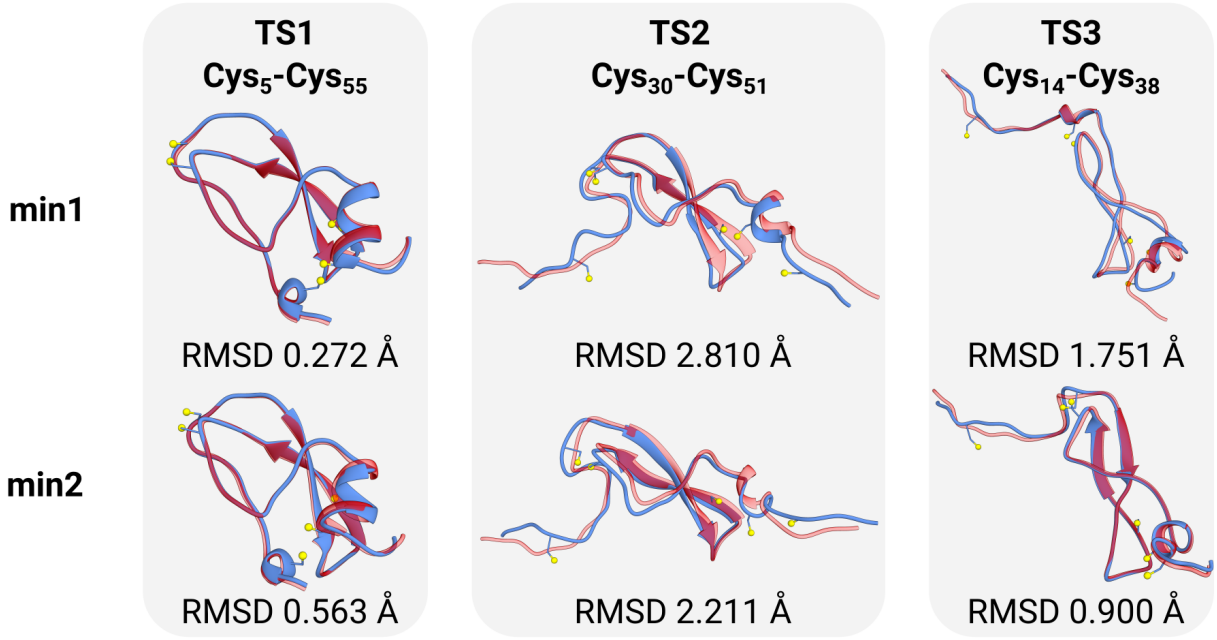

Figure 6: Comparison between the MC-ONIOM3 (solid blue) and GFN-FF (transparent red) reoptimised structures for the minima connected by transition states TS1, TS2, and TS3. "min1" denotes structures with intact disulphide bonds, and "min2" denotes structures with cleaved disulphide bonds. Heavy-atom RMSDs are given below each structure pair. **Positions for Cys-sulfur atoms have been marked for the MC-ONIOM3 geometries.**

tent across the different levels of theory and provide valid insight into the relative stabilities. The same is true for the minima linked by TS3, where residues are sufficiently separated, diminishing the influence of non-covalent interactions and leading to relatively small RMSDs. Only in structures that remain relatively coiled, such as for TS2, do numerous non-covalent interactions persist even after the breaking of two disulphide bonds. Here, the precision of the method to describe the non-covalent interactions significantly influences structural outcomes, resulting in substantial conformational differences (and high RMSDs) between the force-field method and the ONIOM setup. Truncation of non-covalent interactions in the ONIOM layer can be an important source of error in this case,[32,33] as was also seen for cystine in §3.1. Caution is advised when interpreting differences in stability for structures that differ significantly in terms of the non-covalent terms. The mismatches observed in the calculated stabilities of TS1/TS2 using MC-ONIOM3 ($\Delta\Delta G$ of 32.6 kcal mol^{-1}) versus the

force-field ($\Delta\Delta G$ of 9.6 kcal mol^{-1}) probably stem from this factor.

Finally, Fig. 7 illustrates the temperature dependence of the reaction free energy ($\Delta G^{(T)}$) for the three disulphide bond cleavages, calculated with MC-ONIOM3. This graphical representation provides valuable insights into the thermodynamic behaviour of the reactions across a temperature range. The overall positive $\Delta G^{298.15}$ values across all three reactions

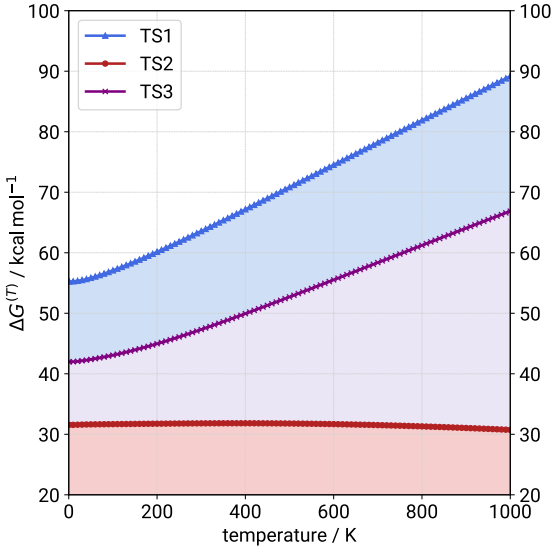

Figure 7: Reaction free energy ($\Delta G^{(T)}$) temperature dependence shown for the three disulphide bond cleavages of TS1 (Cys_5-Cys_{55}), TS2 (Cys_{30}-Cys_{51}), and TS3 (Cys_{14}-Cys_{38}).

show the disulphide bond cleavage to be an endergonic process via the assumed oxidative reaction, as expected. For this type of reaction in the gas phase, a further increase of the temperature is expected to increase the Gibbs free energy $\Delta G^{(T)}$, i.e. making the reaction even less likely to occur at higher temperatures. This trend was confirmed for TS1 and TS3, but surprisingly, TS2 instead exhibits an almost constant, or slightly decreasing, reaction free energy with rising temperature. Identifying a cause for this observation is non-trivial. However, examination of the individual contributions to $\Delta\delta G_{\mathrm{mRRHO}}$ reveals a dominance of (vibrational) entropy with increasing temperature within the reaction free energies shown in Fig. 7. Apparently, the two minima connected by TS2 have similar conformational flexibility, which leads to a stabilisation of the reaction due to the respective entropy contributions. In the case of TS1 and TS3, the minima that contain intact disulphide bonds exhibit more

low-frequency modes contributing significantly to the entropy than the connected minima with broken bonds. This result is consistent with chemical intuition and the expectation that a greater number of such modes should be present for a more rigid structure. We note, however, that in this representation a part of the entropy, namely the landscape entropy associated with other minima, is missing because wider sampling is out of scope for the single-structure supramolecular approach employed for a system of this size.[58] The final free energies might be affected by this omission, although we believe that it would not change the interpretation of the relative stabilities.

4 Conclusion

In this study, we employed the well-known ONIOM methodology to uncover the relative stability of disulphide bonds in the BPTI folding pathway and their associated energetics. A new standalone Fortran library called lwONIOM was introduced to enable multi-layer and -centre ONIOM calculations, exploiting the all-atom general force-field GFN-FF, the semiempirical electronic structure method GFN2-xTB, and calculations at the r^{2}SCAN-3c level of DFT. We provided insights into the computational strengths and limitations of these methods, in particular investigating the robustness of GFN-FF in capturing structural changes and the ability to describe disulphide bond ruptures.

An initial exploration into the oxidative formation of cystine and the three disulphide bond cleavages encountered in BPTI revealed that GFN-FF does not accurately capture the required energetics. The reaction energies for the disulphide bonds, assessed at the MC-ONIOM3 level, exhibit a substantial stabilizing impact on the overall BPTI folding pathway, ranging from 34.4 to 55.6 kcal mol^{-1}. Correspondingly, the associated Gibbs free energies, ranging from 30.9 to 63.6 kcal mol^{-1}, highlight the significant influence of thermostatistical contributions and their non-negligible effect at finite temperatures. **These stabilities can be interpreted in agreement with experimental observations,[75] although a true**

comparison will require to modell alternative Cys-Cys residue combinations and pathways. Description of the latter is currently not within the technical capabilities of GFN-FF but we plan to address this in future research. GFN-FF not only fails to accurately reproduce the relative stabilities of BPTI intermediates but also exhibits an incorrect sign. Intriguingly, a deliberate depiction of disulphide bond ruptures as homolytic dissociation, although chemically incorrect, manages to qualitatively reproduce the high-level energetics. This observation provides a good argument for utilising force-field methods to survey folding pathways, subject to obvious caveats. **The application of implicit solvation expectantly attenuates non-covalent interactions, but does not significantly affect the interpretation of relative disulphide stabilities.**

In conclusion, we provide a qualitatively well-defined reference for the stability of disulphide bonds in biomolecular folding pathways**.** ~~**, highlighting the interdisciplinary nature of computational biology.**~~ In future work, we aim to advance the computational methodology by directly running pathways and calculating transition states at the MC-ONIOM level. While the current project leveraged GFN-FF interfaced in the OPTIM program, upcoming work will use a more direct implementation of the lwONIOM library. In this context, the investigation of (implicit) solvation effects, and the integration of electrostatic rather than purely mechanical embedding in ONIOM is planned. To broaden the scope of our investigations and enable the simulation of even larger systems with higher accuracy, we are planning to update our library further by interfacing coarse-grained potentials, such as the UNited RESidue (UNRES) force-field,[78,79] also recently interfaced to OPTIM.[21] This integration will allow the use of composite QM/SQM/MM/CG methodologies. As we extend the capabilities of our computational tools, we hope that these advances will produce further insight into the dynamics of complex biomolecular systems.

Acknowledgement

PAW acknowledges the Engineering and Physical Sciences Research Council (EPSRC) for funding his studentship through Doctoral Training Partnership EP/W524633/1. PP gratefully acknowledges support by the Alexander von Humboldt Foundation for a Feodor Lynen Research Fellowship.

References

(1) Lensink, M. F.; Brysbaert, G.; Mauri, T.; Nadzirin, N.; Velankar, S.; Chaleil, R. A.; Clarence, T.; Bates, P. A.; Kong, R.; Liu, B., et al. Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment. *Proteins* **2021**, *89*, 1800–1823.

(2) Antoniak, A.; Biskupek, I.; Bojarski, K. K.; Czaplewski, C.; Giełdoń, A.; Kogut, M.; Kogut, M. M.; Krupa, P.; Lipska, A. G.; Liwo, A., et al. Modeling protein structures with the coarse-grained UNRES force field in the CASP14 experiment. *J. Mol. Graph. Model.* **2021**, *108*, 108008.

(3) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A., et al. Applying and improving AlphaFold at CASP14. *Proteins* **2021**, *89*, 1711–1721.

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

(5) Houk, K. N.; Liu, F. Holy Grails for Computational Organic Chemistry and Biochemistry. *Acc. Chem. Res.* **2017**, *50*, 539–543.

(6) Wales, D. J. Exploring Energy Landscapes. *Annu. Rev. Phys. Chem.* **2018**, *69*, 401–425.

(7) Joseph, J. A.; Röder, K.; Chakraborty, D.; Mantell, R. G.; Wales, D. J. Exploring biomolecular energy landscapes. *Chem. Commun.* **2017**, *53*, 6974–6988.

(8) Wales, D. J. Decoding the Energy Landscape: Extracting Structure, Dynamics and Thermodynamics. *Phil. Trans. Roy. Soc. A* **2012**, *370*, 2877–2899.

(9) Klenin, K.; Strodel, B.; Wales, D. J.; Wenzel, W. Modelling Proteins: Conformational Sampling and Reconstruction of Folding Kinetics. *Biochim. Biophys. Acta* **2011**, *1814*, 977–1000.

(10) Wales, D. J.; Bogdan, T. V. Potential Energy and Free Energy Landscapes. *J. Phys. Chem. B* **2006**, *110*, 20765–20776.

(11) Wlodawer, A.; Walter, J.; Huber, R.; Sjölin, L. Structure of bovine pancreatic trypsin inhibitor: Results of joint neutron and X-ray refinement of crystal form II. *J. Mol. Bio.* **1984**, *180*, 301–329.

(12) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y., et al. Atomic-level characterization of the structural dynamics of proteins. *Science* **2010**, *330*, 341–346.

(13) Weissman, J. S.; Kim, P. S. Reexamination of the folding of BPTI: predominance of native intermediates. *Science* **1991**, *253*, 1386–1393.

(14) Creighton, T. E. The disulfide folding pathway of BPTI. *Science* **1992**, *256*, 111–114.

(15) Qin, M.; Wang, W.; Thirumalai, D. Protein folding guides disulfide bond formation. *Proc. Nat. Acad. Sci.* **2015**, *112*, 11241–11246.

(16) Roterman, I.; Sieradzan, A.; Stapor, K.; Fabian, P.; Wesołowski, P.; Konieczny, L. On the need to introduce environmental characteristics in ab initio protein structure prediction using a coarse-grained UNRES force field. *J. Mol. Graph. Model.* **2022**, *114*, 108166.

(17) Lipska, A. G.; Antoniak, A. M.; Wesołowski, P.; Warszawski, A.; Samsonov, S. A.; Sieradzan, A. K. Coarse-grained modeling of the calcium, sodium, magnesium and potassium cations interacting with proteins. *J. Mol. Model.* **2022**, *28*, 201.

(18) Sieradzan, A. K.; Czaplewski, C. R.; Lubecka, E. A.; Lipska, A. G.; Karczynska, A. S.; Gieldon, A. P.; Slusarz, R.; Makowski, M.; Krupa, P.; Kogut, M., et al. Extension of the UNRES Package for Physics-Based Coarse-Grained Simulations of Proteins and Protein Complexes to Very Large Systems. *Biophys. J.* **2021**, *120*, 83a–84a.

(19) Bursch, M.; Mewes, J.-M.; Hansen, A.; Grimme, S. Best-Practice DFT Protocols for Basic Molecular Computational Chemistry**. *Angew. Chem. Int. Ed.* **2022**, *61*, e202205735.

(20) Grimme, S.; Bohle, F.; Hansen, A.; Pracht, P.; Spicher, S.; Stahn, M. Efficient Quantum Chemical Calculation of Structure Ensembles and Free Energies for Nonrigid Molecules. *J. Phys. Chem. A* **2021**, *125*, 4039–4054.

(21) Wesołowski, P. A.; Sieradzan, A. K.; Winnicki, M. J.; Morgan, J. W.; Wales, D. J. Energy landscapes for proteins described by the UNRES coarse-grained potential. *Biophys. Chem.* **2023**, *303*, 107107.

(22) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555.

(23) Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R. Machine Learning Force Fields. *Chem. Rev.* **2021**, *121*, 10142–10186.

(24) Christensen, A. S.; Kubař, T.; Cui, Q.; Elstner, M. Semiempirical Quantum Mechanical Methods for Noncovalent Interactions for Chemical and Biochemical Applications. *Chem. Rev.* **2016**, *116*, 5301–5337.

(25) Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. Extended tight-binding quantum chemistry methods. *WIREs Comput. Mol. Sci.* **2020**, e01493.

(26) Senn, H. M.; Thiel, W. QM/MM Methods for Biomolecular Systems. *Ang. Chem. Int. Ed.* **2009**, *48*, 1198–1229.

(27) Shaik, S.; Cohen, S.; Wang, Y.; Chen, H.; Kumar, D.; Thiel, W. P450 Enzymes: Their Structure, Reactivity, and Selectivity—Modeled by QM/MM Calculations. *Chem. Rev.* **2010**, *110*, 949–1017.

(28) Ahmadi, S.; Barrios Herrera, L.; Chehelamirani, M.; Hostaš, J.; Jalife, S.; Salahub, D. R. Multiscale modeling of enzymes: QM-cluster, QM/MM, and QM/MM/MD: A tutorial review. *Int. J. Quant. Chem.* **2018**, *118*, e25558.

(29) Csizi, K.-S.; Reiher, M. Universal QM/MM approaches for general nanoscale applications. *WIREs Comput. Mol. Sci.* **2023**, *13*, e1656.

(30) Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K. ONIOM: A Multilayered Integrated MO + MM Method for Geometry Optimizations and Single Point Energy Predictions. A Test for Diels-Alder Reactions and Pt(P(t-Bu)3)2 + H2 Oxidative Addition. *J. Phys. Chem.* **1996**, *100*, 19357–19363.

(31) Dapprich, S.; Komáromi, I.; Byun, K.; Morokuma, K.; Frisch, M. J. A new ONIOM implementation in Gaussian98. Part I. The calculation of energies, gradients, vibrational frequencies and electric field derivatives. *J. Mol. Struct: THEOCHEM* **1999**, *461-462*, 1–21.

(32) Chung, L. W.; Sameera, W. M. C.; Ramozzi, R.; Page, A. J.; Hatanaka, M.; Petrova, G. P.; Harris, T. V.; Li, X.; Ke, Z.; Liu, F.; Li, H.-B.; Ding, L.; Morokuma, K. The ONIOM Method and Its Applications. *Chem. Rev.* **2015**, *115*, 5678–5796.

(33) Chung, L. W.; Hirao, H.; Li, X.; Morokuma, K. The ONIOM method: its foundation and applications to metalloenzymes and photobiology. *WIREs Comput. Mol. Sci.* **2012**, *2*, 327–350.

(34) Seeber, P.; Seidenath, S.; Steinmetzer, J.; Gräfe, S. Growing Spicy ONIOMs: Extending and generalizing concepts of ONIOM and many body expansions. *WIREs Comput. Mol. Sci.* **2023**, *13*, e1644.

(35) Ufimtsev, I. S.; Luehr, N.; Martinez, T. J. Charge Transfer and Polarization in Solvated Proteins from Ab Initio Molecular Dynamics. *J. Phys. Chem. Lett.* **2011**, *2*, 1789–1793.

(36) Kamerlin, S. C. L.; Haranczyk, M.; Warshel, A. Progress in Ab Initio QM/MM Free-Energy Simulations of Electrostatic Energies in Proteins: Accelerated QM/MM Studies of pKa, Redox Reactions and Solvation Free Energies. *J. Phys. Chem. B* **2009**, *113*, 1253–1272.

(37) Duarte, F.; Amrein, B. A.; Blaha-Nelson, D.; Kamerlin, S. C. Recent advances in QM/MM free energy calculations using reference potentials. *Biochim. Biophys. Acta* **2015**, *1850*, 954–965.

(38) Spicher, S.; Grimme, S. Robust Atomistic Modeling of Materials, Organometallic, and Biochemical Systems. *Angew. Chem. Int. Ed.* **2020**, *132*, 15795–15803.

(39) Furness, J. W.; Kaplan, A. D.; Ning, J.; Perdew, J. P.; Sun, J. Accurate and Numerically Efficient r^{2}SCAN Meta-Generalized Gradient Approximation. *J. Phys. Chem. Lett.* **2020**, *11*, 8208–8215.

(40) Grimme, S.; Hansen, A.; Ehlert, S.; Mewes, J.-M. r^{2}SCAN-3c: A "Swiss army knife" composite electronic-structure method. *J. Chem. Phys.* **2021**, *154*, 064103.

(41) Cui, Q.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Karplus, M. A QM/MM Implemen-

tation of the Self-Consistent Charge Density Functional Tight Binding (SCC-DFTB) Method. *J. Phys. Chem. B* **2001**, *105*, 569–585.

(42) Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.

(43) Pracht, P.; Morgan, J. W. R.; Wales, D. J. Exploring energy landscapes for solid-state systems with variable cells at the extended tight-binding level. *J. Chem. Phys.* **2023**, 064801.

(44) Furman, D.; Naumkin, F.; Wales, D. J. Energy Landscapes of Carbon Clusters from Tight-Binding Quantum Potentials. *J. Phys. Chem. A* **2022**, *126*, 2342–2352.

(45) Gorges, J.; Grimme, S.; Hansen, A.; Pracht, P. Towards understanding solvation effects on the conformational entropy of non-rigid molecules. *Phys. Chem. Chem. Phys.* **2022**, *24*, 12249–12259.

(46) Spicher, S.; Grimme, S. Efficient Computation of Free Energy Contributions for Association Reactions of Large Molecules. *J. Phys. Chem. Lett.* **2020**, *11*, 6606–6611.

(47) Plett, C.; Katbashev, A.; Ehlert, S.; Grimme, S.; Bursch, M. ONIOM meets xtb: efficient, accurate, and robust multi-layer simulations across the periodic table. *Phys. Chem. Chem. Phys.* **2023**, *25*, 17860–17868.

(48) Pyykkö, P.; Atsumi, M. Molecular Single-Bond Covalent Radii for Elements 1–118. *Chem. Eur. J. 15*, 186–197.

(49) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB – An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.

(50) lwONIOM: A light-weight multi-center n-level ONIOM interface. `https://github.com/crest-lab/lwoniom`, February 22, 2024.

(51) Pracht, P.; Grimme, S.; Bannwarth, C.; Bohle, F.; Ehlert, S.; Feldmann, G.; Gorges, J.; Müller, M.; Neudecker, T.; Plett, C.; Spicher, S.; Steinbach, P.; Wesołowski, P. A.; Zeller, F. CREST - A program for the exploration of low-energy molecular chemical space. *manuscript in preparation*

(52) Conformer-Rotamer Ensemble Sampling Tool (CREST). `https://github.com/crest-lab/crest`, February 22, 2024.

(53) Neese, F. The ORCA program system. *WIREs Comput. Mol. Sci.* **2012**, *2*, 73–78.

(54) Neese, F.; Wennmohs, F.; Becker, U.; Riplinger, C. The ORCA quantum chemistry program package. *J. Chem. Phys.* **2020**, *152*, 224108.

(55) Mardirossian, N.; Head-Gordon, M. ωB97X-V: A 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy. *Phys. Chem. Chem. Phys.* **2014**, *16*, 9904–9924.

(56) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

(57) Grimme, S. Supramolecular binding thermodynamics by dispersion corrected density functional theory. *Chem. Eur. J.* **2012**, *18*, 9955–9964.

(58) Pracht, P.; Grimme, S. Calculation of absolute molecular entropies and heat capacities made simple. *Chem. Sci.* **2021**, *12*, 6551–6568.

(59) `OPTIM`: A program for geometry optimisation and pathway calculations. http://www-wales.ch.cam.ac.uk/software.html.

(60) Röder, K.; Joseph, J. A.; Husic, B. E.; Wales, D. J. Energy Landscapes for Proteins: From Single Funnels to Multifunctional Systems. *Adv. Theory Simul.* **2019**, *2*, 1800175.

(61) Trygubenko, S. A.; Wales, D. J. A Doubly Nudged Elastic Band Method for Finding Transition States. *J. Chem. Phys.* **2004**, *120*, 2082–2094.

(62) Sheppard, D.; Terrell, R.; Henkelman, G. Optimization methods for finding minimum energy paths. *J. Chem. Phys.* **2008**, *128*, 134106.

(63) Mills, G.; Jónsson, H.; Schenter, G. K. Reversible work transition state theory: application to dissociative adsorption of hydrogen. *Surface Science* **1995**, *324*, 305–337.

(64) Jónsson, H.; Mills, G.; Jacobsen, K. W. In *Classical and Quantum Dynamics in Condensed Phase Simulations*; Berne, B. J., Ciccotti, G., Coker, D. F., Eds.; World Scientific: Singapore, 1998; Chapter 16, pp 385–404.

(65) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.* **2000**, *113*, 9901–9904.

(66) Henkelman, G.; Jónsson, H. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.* **2000**, *113*, 9978–9985.

(67) Wales, D. J.; Carr, J. M. Quasi-Continuous Interpolation Scheme for Pathways between Distant Configurations. *J. Chem. Theory Comput.* **2012**, *8*, 5020–5034.

(68) Röder, K.; Wales, D. J. Predicting Pathways between Distant Configurations for Biomolecules. *J. Chem. Theory Comput.* **2018**, *14*, 4271–4278.

(69) Ehlert, S.; Stahn, M.; Spicher, S.; Grimme, S. Robust and Efficient Implicit Solvation Model for Fast Semiempirical Methods. *J. Chem. Theory Comput.* **2021**, *17*, 4250–4261.

(70) Flores-Huerta, A. G.; Tkatchenko, A.; Galván, M. Nature of Hydrogen Bonds and S···S Interactions in the l-Cystine Crystal. *J. Phys. Chem. A* **2016**, *120*, 4223–4230.

(71) Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S. A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Phys. Chem. Chem. Phys.* **2017**, *19*, 32184–32215.

(72) Najibi, A.; Goerigk, L. The Nonlocal Kernel in van der Waals Density Functionals as an Additive Correction: An Extensive Analysis with Special Emphasis on the B97M-V and ωB97M-V Approaches. *J. Chem. Theory Comput.* **2018**, *14*, 5725–5738.

(73) Becker, O. M.; Karplus, M. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *J. Chem. Phys.* **1997**, *106*, 1495–1517.

(74) Wales, D. J.; Miller, M. A.; Walsh, T. R. Archetypal energy landscapes. *Nature* **1998**, *394*, 758–760.

(75) Darby, N. J.; Morin, P. E.; Talbo, G.; Creighton, T. E. Refolding of bovine pancreatic trypsin inhibitor via non-native disulphide intermediates. *J. Mol. Biol.* **1995**, *249*, 463–477.

(76) Weissman, J. S.; Kim, P. S. Kinetic role of nonnative species in the folding of bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci.* **1992**, *89*, 9900–9904.

(77) Hatahet, F.; Ruddock, L. W. Protein disulfide isomerase: a critical evaluation of its function in disulfide bond formation. *Antioxid Redox Signal* **2009**, *11*, 2807–2850.

(78) Czaplewski, C.; Karczyńska, A.; Sieradzan, A. K.; Liwo, A. UNRES server for physics-based coarse-grained simulations and prediction of protein structure, dynamics and thermodynamics. *Nucleic Acids Res.* **2018**, *46*, W304–W309.

(79) Liwo, A.; Czaplewski, C.; Sieradzan, A. K.; Lipska, A. G.; Samsonov, S. A.; Murarka, R. K. Theory and practice of coarse-grained molecular dynamics of biologically important systems. *Biomolecules* **2021**, *11*, 1347.

Graphical TOC Entry