# Mutational Basin-Hopping – Combined Structure and Sequence Optimisation for Biomolecules

Konstantin Röder and David J. Wales*

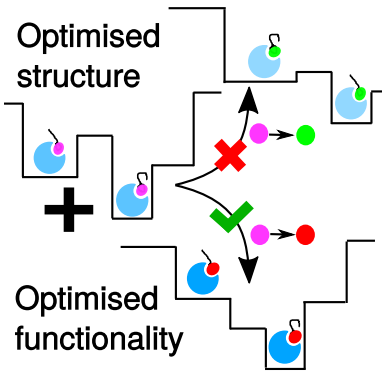
*Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, UK.*

E-mail: dw34@cam.ac.uk

*To whom correspondence should be addressed

**Abstract**

*The study of energy landscapes has led to a good understanding of how and why proteins and nucleic acids adopt their native structure. Through evolution, sequences have adapted until they exhibit a strongly funnelled energy landscape, stabilising the native fold. Design of artificial biomolecules faces the challenge of creating similar stable, minimally frustrated and functional sequences. Here we present a biminimisation approach, mutational basin-hopping, in which we simultaneously use global optimisation to optimise the energy and a target function describing a desired property of the system. This optimisation of structure and sequence is a generalised basin-hopping method, and produces an efficient design process, which can target properties such as binding affinity or solubility.*




Folding of evolved proteins found in nature is based on a geometrically biased energy landscape with a single[1–3] or multiple[4] funnels, such that a dominant structural ensemble, the native state, is adopted at equilibrium. While evolution over a long time scale has managed to mostly exclude undesired features in the landscape,[5,6] artificially designed sequences are likely to exhibit higher frustration, corresponding to competing low-lying morphologies separated by high energy barriers. In a designed system, the target state is less likely to optimise all the packing and interaction requirements, which is necessary to satisfy the principle of minimal frustration.[1] While the residual frustration can lead to multifunctional molecules,[7] the design process can easily produce highly frustrated sequences, without well-defined native structure. Furthermore,

the large number of possible sequences prohibits extensive experimental screening,[8] and also complicating simulations. In fact, the optimisation of a sequence to adopt a specific backbone fold is NP hard,[9] and the algorithmic costs are therefore high. Nonetheless, significant progress has been made,[10] either starting from known folds and altering the sequence, or for entirely new sequences and novel folds,[11–13] for example in Top7.[14] A key step in the current state-of-the-art methods is the separation into the selection of a target structure and the subsequent search for a sequence supporting this structure,[15] known as the inverse protein folding problem.[16] Such an approach may take frustration into account, for example using a fitness function.[15]

However, even if a specific backbone fold is designed successfully, it remains an open question whether the desired properties of the new sequence are indeed observed by experiment. A common approach is therefore the screening of a large number of components, either in simulation or experiment. This approach has yielded some remarkable results, for example in the engineering of ion channels,[17] nanopores,[18] the design of antibodies,[19] and ion-binding scaffolds;[20] but the costs for such screening processes is high.

Instead of optimising the sequence to match a specific backbone fold, we suggest a new scheme, mutational basin-hopping, where global optimisation techniques are applied to the energy of a given sequence, to predict the most favourable structure of a given fold, combined with a second penalty function, which is based on an observable property of interest. Such a biminimisation has previously been applied to other problems, in particular to determine the optimal size of nanoparticles,[21] to probe the compositions of nanoalloys,[22] and to study to energy and chemical potential variations simultaneously.[23] Here we present the algorithm in the context of biomolecular optimisation for the first time, with a discussion of some key properties, and provide results for a proof of concept simulation.

Mutational basin-hopping global optimisation is a generalised basin-hopping technique.[21,22] For an all-atom force field, mutations change the number of atoms as well as the element for some atoms. Mutations therefore correspond to grand canonical

and alchemical transformations of the system. Within the configuration space for each sequence the optimisation must address the overall potential energy to locate the lowest energy structure given a sequence. Between sequences, a second penalty function is necessary to accept or reject mutations, but this function depends on the system and the target property. Generalised basin-hopping[21,22] employs steps in a second metric space, in this case the sequence space, over relatively large intervals of $N$ steps, allowing for location of low energy structures for every sequence. When a step in the second metric space is taken, the system is allowed some exploration of the landscape for the new sequence, with $n$ steps before the mutation is accepted or rejected. In outline, the algorithm proceeds as follows:

1. Basin-hopping global optimisation searches for the lowest energy configuration. For every structure, $\mathbf{X}$, the penalty function $f_{\mathrm{mut}}(\mathbf{X})$ is computed. The lowest value, $f_{\mathrm{min}}^{\mathrm{prev}}$, is saved as the structure search progresses. Here, $\mathbf{X}$ is a vector of Cartesian atomic coordinates for a local minimum, which is located after energy minimisation following a propagation move. We note that $f_{\mathrm{mut}}$ can be maximised rather than minimised (as for example in Fig. 1), and we can simply change the sign of $f_{\mathrm{mut}}$ appropriately.

2. A residue is chosen for mutation, which can be limited to a particular set of amino or nucleic acids, or include all possibilities. The subsequent choice of the site and the mutation, as well as the possible limitations on the allowed mutations, allows for nuanced changes, for example in a binding pocket or a hydrophobic patch.

3. After the data for the old sequence is saved, the new sequence is initialised, and all related local properties, such as rigidification[24,25] and group rotation setup,[26,27] are completed, the structure search continues with the new sequence. Again, the lowest value of the penalty function, $f_{\mathrm{min}}^{\mathrm{new}}$, is saved.

4. After $n$ steps, an accept/reject criterion is applied using $f_{\mathrm{min}}^{\mathrm{prev}}$ and $f_{\mathrm{min}}^{\mathrm{new}}$ to score the new sequence. If the new sequence is accepted, the search is continued, otherwise the old sequence is restored and the search restarts from the last sampled

configuration fir that system.

An illustration of the algorithm is given in Fig. 1.

While a number of features are identical to those used in basin-hopping[28] and grand-canonical basin-hopping,[23] two aspects merit further discussion: the computation of the new coordinates as the number and identity of atoms change, and the penalty functions used in different cases.

Mutations between different amino or nucleic acids require changes in atomic identity and the number of atoms. While the information about the change in identity, such as bond strengths and charges, is provided by force field libraries, the change in the number of atoms and the necessary creation of coordinates for the atoms in the new sequence require additional treatment. We can use standard orientations, as provided in force field or rotamer libraries, but it is desirable to preserve as much of the original structure as possible to reduce the likelihood of atom clashes and unfavourable interactions. The only changes applied are in the side chains and nucleobases, respectively. Atom positions are conserved for atoms of the same element and hybridisation state before and after the mutation. All other positions are created based on hybridisation states and chirality constraints. This procedure proved efficient in creating relatively low energy starting structures after mutations were proposed.

The ability to optimise sequences requires the definition of a function that describes the target properties of the biomolecule. A key condition is that it must be possible to calculate this property from the molecular coordinates and the computational cost to estimate this function should be as low as possible.

As most, if not all, biomolecules are only able to fulfil their biological function in combination with other species, binding energies are important. For example, the recognition of small peptides by their carrier and target proteins, binding of antibodies to target epitopes, and the interaction between regulatory nucleic acids and proteins all fall into this category. To estimate binding affinities, we employ all the polar interactions contributing to the energy difference between the bound ligand and target, and the unbound state. A simple way to treat the unbound state is to locate the centres

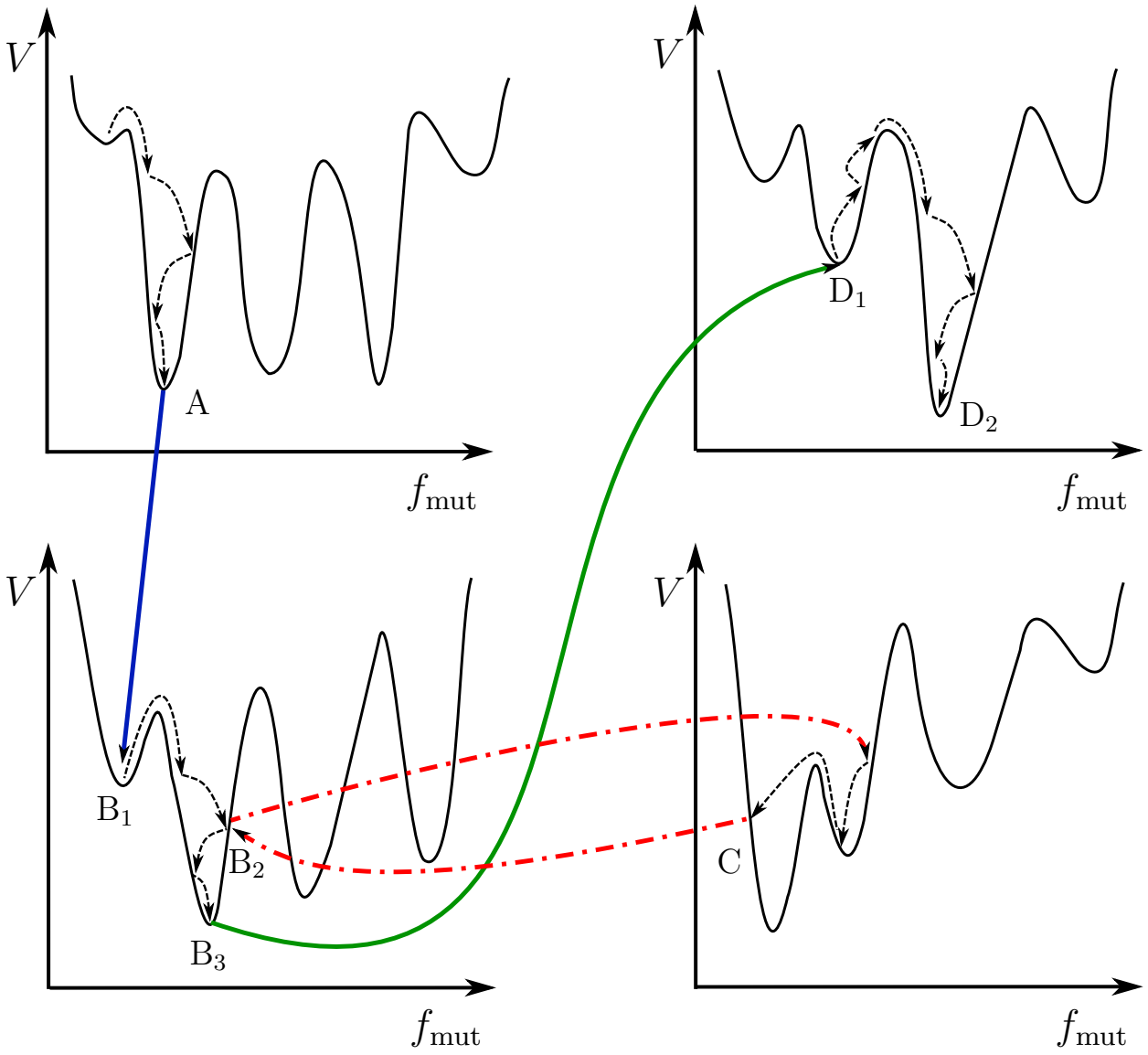



Figure 1: Mutational basin-hopping: In the four schematic landscapes the black arrows represent basin-hopping on the individual landscapes corresponding to the sequences A, B, C and D. Mutation from A leads to state $B_1$ (blue). This mutation is accepted based on the value of $f_{\mathrm{mut}}$ (in this case larger values are targeted). Basin-hopping then propagates the system to state $B_2$. Another mutation leads to sequence C, which is later rejected (red), and basin-hopping continues for mutation B, leading to state $B_3$. The next mutation leads to state $D_1$ (green), and subsequently to state $D_2$. Overall two mutations are accepted.

of mass for separate components in the bound state, define a vector between them, and then translate one molecule along the vector away from the other component. The cost for this approach is minimal, as it only requires the computation of the centres of mass, an addition of two arrays, and one call to an energy routine. This approach yields a useful penalty function for sequence optimisation. At the same time the exploration of sequence space might be used to better understand the binding interactions and pattern recognition within the complex, providing insight into the binding mechanism at an atomistic level of detail.

Another important property of biomolecules is their solubility. The solvation energy is readily available in most biomolecular potentials for all conformations, as it is calculated to determine the total energy. Various other options exist for the penalty function. For example, we could optimise the sequence to match a specified backbone configuration, providing a connection to other methods for protein design.

As proof of principle we have applied mutational basin-hopping global optimisation to the neurophysin II – oxytocin complex. Oxytocin and vasopressin are two nonapeptide hormones, which exhibit a strong dependence of binding affinity to carrier and target proteins with respect to sequence mutations.[29–36] The two hormones are related by two mutations, namely I3F and L8R, from oxytocin to vasopressin. The structural ensembles of the two hormones have been studied with a number of experimental and computational methods.[4,37–42] Neurophysin II is the carrier protein for vasopressin, but can bind oxytocin as well. The two hormones provide a well-studied system and exhibit a remarkable specificity in their binding affinity, a desirable target for designed peptides.

Our objective here is intended to demonstrate mutational basin-hopping, rather than to provide an extensive study of the bound complex. The starting point for the simulation was the crystal structure[43] and mutations were allowed in positions 2, 3 and 8, since residue 2 has been associated with a large influence on binding,[38,42] and residues 3 and 8 permit the switch between oxytocin and vasopressin.The simulation was 150,000 basin-hopping steps long and mutations were accepted or rejected after

4,000 steps, using the interaction energy between the carrier and the hormone in the penalty function. The separation imposed between hormone and carrier to calculate the penalty function was 150 Å. Exploration of the configuration space used group rotation moves,[24,25] and cis-trans isomers and chirality were conserved.

The lowest potential energy and the interaction energy for the structures encountered are illustrated in Fig. 2. The simulation features an early set of accepted mutations (L8R and I3F), which transform the hormone into vasopressin. We expect vasopressin to be a better ligand for neurophysin II, and the algorithm correctly identifies this property. Later in the run there are many proposed mutations of residue 8, back to the oxytocin sequence, all of which are rejected.

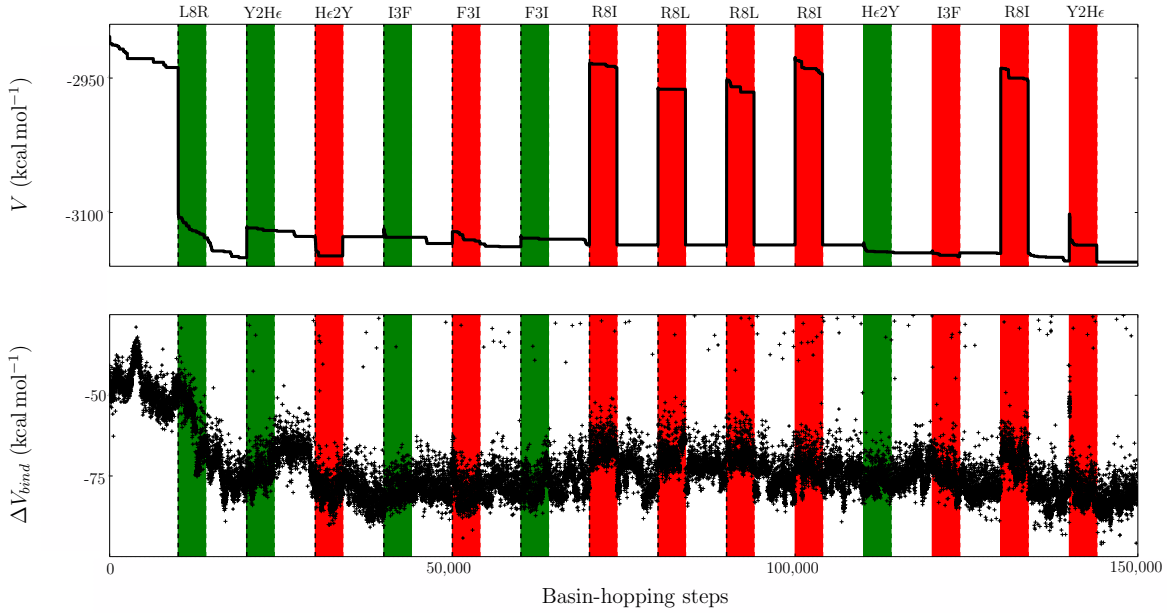



Figure 2: Progress of a mutational basin-hopping run of 150,000 steps (local minimisations), with 10,000 steps between each attempted mutation, and 4,000 steps before a mutation is accepted or rejected. The top panel shows the lowest potential energy encountered for the current sequence. The bottom panel shows the binding interaction between the hormone and the carrier protein. The plotted values for the binding energies correspond to 105,986 structures with physically reasonable energies. Accepted mutations are highlighted in green, and rejected mutations are in red.

In summary, we have developed a new algorithm, mutational basin-hopping, which involves biminimisation[21] of sequence and structure. The algorithm is flexible, accommodating a wide variety of potentials, propagation moves, and penalty functions.

Application to the neurophysin II – oxytocin complex as proof of principle shows that the algorithm can faithfully reproduce experiment. Pierce and Winfree hypothesised that an approach looking beyond the backbone configuration might provide practical and computational improvements over other methods.[9] The clear connection to energy landscape theory in the present work provides this advance.

## Acknowledgement

The authors acknowledge funding from the Engineering and Physical Sciences Research Council UK.

## References

(1) Bryngelson, J. D.; Wolynes, P. G. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 7524–7528.

(2) Leopold, P. E.; Montal, M.; Onuchic, J. N. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 8721–8725.

(3) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins* **1995**, *21*, 167–195.

(4) Röder, K.; Wales, D. J. Evolved minimal frustration in multifunctional biomolecules. *J. Phys. Chem. B* **2018**, in press.

(5) Morcos, F.; Schafer, N. P.; Cheng, R. R.; Onuchic, J. N.; Wolynes, P. G. Co-evolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proc. Natl Acad. Sci. USA* **2014**, *111*, 12408–12413.

(6) Wolynes, P. G. Evolution, energy landscapes and the paradoxes of protein folding. *Biochimie* **2015**, *119*, 218–230.

(7) Tyka, M. D.; Keedy, D. A.; André, I.; DiMaio, F.; Song, Y.; Richardson, D. C.; Richardson, J. S.; Baker, D. Alternate states of proteins revealed by detailed energy landscape mapping. *J. Mol. Biol.* **2011**, *405*, 607–618.

(8) Jäckel, C.; Kast, P.; Hilvert, D. Protein Design by Directed Evolution. *Annu. Rev. Biophys.* **2008**, *37*, 153–173.

(9) Pierce, N. A.; Winfree, E. Protein design in NP-hard. *Protein Eng.* **2002**, *15*, 779–782.

(10) Huang, P.-S.; Boyken, S. E.; Baker, D. The coming of age of de novo protein design. *Nature* **2016**, *537*, 320–327.

(11) Kim, D. E.; Chivian, D.; Baker, D. Protein structure prediction and analysis using the Rosetta server. *Nucleic Acids Res.* **2004**, *32*, W526–W531.

(12) Das, R.; Baker, D. Macromolecular Modeling with Rosetta. *Annu. Rev. Biochem.* **2008**, *77*, 363–382.

(13) Koga, N.; Tatsumi-Koga, R.; Liu, G.; Xiao, R.; Acton, T. B.; Montelione, G. T.; Baker, D. Principles for designing ideal protein structure. *Nature* **2012**, *491*, 222–227.

(14) Kuhlman, B.; Dantas, G.; Ireton, G. C.; Varani, G.; Stoddard, B. L.; Baker, D. Design of a novel globular protein fold with atomic-level accuracy. *Science* **2003**, *302*, 1364–1368.

(15) Betancourt, M. R.; Thirumalai, D. Protein sequence design by energy landscaping. *J. Chem. Phys. B* **2002**, *106*, 599–609.

(16) Drexler, K. E. Molecular engineering: An approach to the development of general capabilities for molecular manipulation. *Proc. Natl. Acad. Sci. USA* **1981**, *78*, 5275–5278.

(17) Lella, M.; Mahalashmi, R. Engineering a transmembrane nanopore ion channel from a membrane breaker peptide. *J. Chem. Phys. Lett.* **2016**, *7*, 2298–2303.

(18) Liu, A.; Zhao, Q.; Krishantha, D. M.; Guan, X. Unzipping of double-stranded DNA in engineered alpha-hemolysin pores. *J. Chem. Phys. Lett.* **2011**, *2*, 1372–1376.

(19) Greene, D.; Po, T.; Pan, J.; Tabibian, T.; Luo, R. Computational analysis for the rational design of anti-amyloid beta (A$\beta$) antibodies. *J. Chem. Phys. B* **2018**, *122*, 4521–4536.

(20) Jonnalagadda, S. V. R.; Kokotidou, C.; Orr, A. A.; Fotopoulou, E.; Henderson, K. J.; Choi, C. H.; Lim, W. T.; Choi, S. J.; Jeong, H. K.; Mitraki, A. et al. Computational design of functional amyloid materials with cesium binding, deposition and capture properties. *J. Chem. Phys. B* **2018**, *122*, 7555–7568.

(21) Schebarchov, D.; Wales, D. J. Structure prediction for multicomponent materials using biminima. *Phys. Rev. Lett.* **2014**, *113*, 156102.

(22) Schebarchov, D.; Wales, D. J. Quasi-combinatorial energy landscapes for nanoalloy structure optimisation. *Phys. Chem. Chem. Phys.* **2015**, *12*, 902–909.

(23) Calvo, F.; Schebarchov, D.; Wales, D. J. Grand and semi-grand canonical basin-hopping. *J. Chem. Theory Comput.* **2015**, *17*, 28331–28338.

(24) Mochizuki, K.; Whittleston, C. S.; Somani, S.; Kusumaatmaja, H.; Wales, D. J. A conformational factorisation approach for estimating the binding free energies of macromolecules. *Phys. Chem. Chem. Phys.* **2014**, *16*, 2842–2853.

(25) Oakley, M. T.; Johnston, R. L. Energy landscapes and global optimization of self-assembling cyclic peptides. *J. Chem. Theory Comput.* **2014**, *10*, 1810–1816.

(26) Kusumaatmaja, H.; Whittleston, C. S.; Wales, D. J. A local rigid body framework for global optimization of biomolecules. *J. Chem. Theory Comput.* **2012**, *8*, 5159–5165.

(27) Rühle, V.; Kusumaatmaja, H.; Chakrabarti, D.; Wales, D. J. Exploring energy landscapes: Metrics, pathways, and normal-mode analysis for rigid-body molecules. *J. Chem. Theory Comput.* **2013**, *9*, 4026–4034.

(28) Wales, D. J.; Doye, J. P. K. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *J. Chem. Phys. A* **1997**, *101*, 5111–5116.

(29) Du Vigneaud, V.; Ressler, C.; Trippett, S. The sequence of amino acids in oxytocin, with a proposal for the structure of oxytocin. *J. Biol. Chem.* **1953**, *205*, 949–957.

(30) Bodansky, M.; Bath, R. J. Hindered amines in peptide synthesis. Synthesis of 7-glycine-oxytocin. *Chem. Commun.* **1968**, 766.

(31) Bespalova, Z.; Martynov, V.; Titov, M. New oxytocin analoguess7-glycine-oxytocin and 7-D-leucine-oxytocin. *Zh. Obs. Khim* **1968**, *38*, 1684–1687.

(32) Walter, R.; Yamanaka, T.; Sakakibarat, S. A neurohypophyseal hormone analog with selective oxytocin-like activities and resistance-to enzymatic inactivation: An approach to the design of peptide drugs. *Proc. Natl. Acad. Sci. USA* **1974**, *71*, 1901–1905.

(33) Lowbridge, J.; Manning, M.; Haldar, J.; Sawyer, W. H. Synthesis and some pharmacological properties of [4-threonine,7-glycine]oxytocin, [1-(L-2-hydroxy-3-mercaptopropanoic acid),4-threonine,7-glycine]oxytocin (hydroxy[thr4, gly7]oxytocin), and [7-glycine]oxytocin, peptides with high oxytocic-antidiuretic selectivity. *J. Med. Chem.* **1977**, *20*, 120–123.

(34) Barberis, C.; Morin, D.; Durroux, T.; Mouillac, B.; Guillon, G.; Seyer, R.; Hibert, M.; Tribollet, E.; Manning, M. Molecular pharmacology of AVP and OT receptors and therapeutic potential. *Drug News Perspect.* **1999**, *12*, 279.

(35) Willcutts, M. D.; Felner, E.; White, P. C. Autosomal recessive familial neurohypophyseal diabetes insipidus with continued secretion of mutant weakly active vasopressin. *Hum. Mol. Genet.* **1999**, *8*, 1303–1307.

(36) Manning, M.; Stoev, S.; Chini, B.; Durroux, T.; Mouillac, B.; Guillon, G. Peptide and non-peptide agonists and antagonists for the vasopressin and oxytocin V1a, V1b, V2 and OT receptors: research tools and potential therapeutic agents. *Prog. Brain Res.* **2008**, *170*, 473–512.

(37) Wu, C. K.; Hu, B.; Rose, J. P.; Liu, Z. J.; Nguyen, T. L.; Zheng, C.; Breslow, E.; Wang, B. C. Structures of an unliganded neurophysin and its vasopressin complex: implications for binding and allosteric mechanisms. *Protein Sci.* **2001**, *8*, e58113.

(38) Rittig, S.; Siggaard, C.; Ozata, M.; Yetkin, I.; Gregersen, N.; Pedersen, E. B.; Robertson, G. L. Autosomal dominant neurohypophyseal diabetes insipidus due to substitution of histidine for tyrosine2 in the vasopressin moiety of the hormone precursor. *J. Clin. Endocrinol. Metab.* **2002**, *87*, 3351–3355.

(39) Slusarz, M. J.; Slusarz, R.; Ciarkowski, J. Molecular dynamics simulation of human neurohypophyseal hormone receptors complexed with oxytocin-modeling of an activated state. *J. Pept. Sci.* **2006**, *12*, 171–179.

(40) Slusarz, M. J.; Giełdon, A.; Slusarz, R.; Ciarkowski, J. Analysis of interactions responsible for vasopressin binding to human neurohypophyseal hormone receptors - molecular dynamics study of the activated receptor-vasopressin-G$\alpha$ systems. *J. Pept. Sci.* **2006**, *12*, 180–189.

(41) Sikorska, E.; Rodziewicz-Motowidło, Conformational studies of vasopressin and mesotocinusing NMR spectroscopy and molecular modelling methods. Part I: Studies in water. *J. Pept. Sci.* **2008**, *14*, 76–84.

(42) Yedvabny, E.; Nerenberg, P. S.; So, C.; Head-Gordon, T. Disordered structural

ensembles of vasopressin and oxytocinand their mutants. *J. Phys. Chem. B* **2015**, *119*, 896–905.

(43) Rose, J. P.; Wu, C. K.; Hsiao, C. D.; Breslow, E.; Wang, B. C. Crystal structure of the neurophysin-oxytocin complex. *Nat. Struct. Biol.* **1996**, *3*, 163–169.