Chemistry–A European Journal

Research Article
doi.org/10.1002/chem.202302375

Chemistry Europe
European Chemical Societies Publishing

www.chemeurj.org

Application of Machine Learning Algorithms to Metadynamics for the Elucidation of the Binding Modes and Free Energy Landscape of Drug/Target Interactions: a Case Study

Gohar Ali Siddiqui+,[a] Julia A. Stebani+,[b] Darren Wragg,[b] Phaedon-Stelios Koutsourelakis,[c] Angela Casini,*[b] and Alessio Gagliardi*[a]

In the context of drug discovery, computational methods were able to accelerate the challenging process of designing and optimizing a new drug candidate. Amongst the possible atomistic simulation approaches, metadynamics (metaD) has proven very powerful. However, the choice of collective variables (CVs) is not trivial for complex systems. To automate the process of CVs identification, two different machine learning algorithms were applied in this study, namely DeepLDA and Autoencoder, to the metaD simulation of a well-researched drug/target complex, consisting in a pharmacologically relevant non-canonical DNA secondary structure (G-quadruplex) and a metallodrug acting as its stabilizer, as well as solvent molecules.

Introduction

Contemporary drug discovery processes start commonly with identifying and validating a biologically relevant target that can be modulated with drug molecules. In this context, computational approaches can accelerate the challenging process of designing and optimizing a new drug candidate.[1] Amongst the possible methods, molecular dynamics (MD) is an essential tool to study phenomena in chemical systems at the atomistic level. This is also true for biological systems, including drug/target interactions,[2] where the temporal and spatial resolution provided by the methods is crucial to understand emergent phenomena at macro scale. However, as soon as the underlying phenomena being studied happens at a time scale longer than a few nanoseconds, the sampling becomes rare.[3] Some tailored computational architectures are able to achieve millisecond scale,[4] but these are special cases and generally achievable timescales remain in the microsecond regime. One of the approaches used to alleviate this limitation is through enhancing the sampling of the configuration space by biasing the simulation.[5] Amongst the enhanced sampling methods, free-energy perturbation,[6] umbrella sampling,[7] replica exchange,[8] metadynamics,[9] steered MD,[10] accelerated MD,[11] milestoning,[12] transition-path sampling,[13] and their many possible combinations, are now widely used. Metadynamics-(metaD)-based methods include a broad family of enhanced sampling techniques enabling fast exploration of the underlying free-energy landscape of rare events. To this aim, a set of order parameters, usually referred to as collective variables (CVs), is used to approximate the actual reaction coordinate of the process.[14] Starting with the work of *Gervasio* et al.,[15] in recent years, metaD approaches have been applied to a number of ligand-target complexes, demonstrating its ability to characterize binding and unbinding paths, to treat conformation flexibility, and to compute free-energy profiles.[16]

Despite the availability of metaD based methods,[9,17] the prerequisite for efficient exploration of CV space is the knowledge of "good" CVs.[5,18] To select a number of hand-picked CVs require deep chemical intuition about the system dynamics and becomes increasingly difficult for complex (biological) systems. To automate this process, a number of procedures have been developed,[19] many of them based on machine learning (ML) approaches.[18a,20] Some methods like Deep Linear Discriminant Analysis (Deep-LDA)[18a] are actually focused on sampling the transitions between different metastable basins rather than finding the ideal CVs. To achieve this, a set of informed CVs are used as input to a non-linear model (e.g. a Neural Network), a lower-dimensional output of which is used in a Linear Discriminant Analysis (LDA) to get maximum discrimination

[a] *G. A. Siddiqui,+ Prof. A. Gagliardi*
Professorship of Simulation of Nanosystems for Energy Conversion
Department of Electrical and Computer Engineering
School of Computation, Information and Technology
Technical University of Munich (TUM)
Hans-Piloty-Str. 1, 85748 Garching b. München (Germany)
E-mail: alessio.gagliardi@tum.de

[b] *J. A. Stebani,+ Dr. D. Wragg, Prof. A. Casini*
Chair of Medicinal and Bioinorganic Chemistry
Department of Chemistry, School of Natural Sciences
Technical University of Munich (TUM)
Lichtenbergstr. 4, 85748 Garching b. München (Germany)
E-mail: angela.casini@tum.de

[c] *Prof. P.-S. Koutsourelakis*
Professorship for Data-driven Materials Modeling
School of Engineering and Design
Technical University of Munich (TUM)
Boltzmannstr. 15, 85748 Garching b. München (Germany)

[+] *These authors contributed equally to this manuscript.*

Supporting information for this article is available on the WWW under https://doi.org/10.1002/chem.202302375

between two known energy minima. The system is then constrained to transit between the known minima, resulting in a dense sampling of the transition states between them, and enabling monitoring the influence of additional effects (e.g. water) on state transitions. In this regard, a prior knowledge of the metastable states and relevant CVs is crucial.

On the other hand, deep neural network architectures like *Autoencoders*[21] have been shown to be able to obtain a good set of CVs[22] using general coordinates of a system. This kind of network is used typically in dimensionality reduction schemes and does not require prior selection of CVs. The training loss is the error in the ability of the network to represent the configurational state of a system (defined by the general coordinates, that can be also basic atomic distances) in a lower dimensional space. This can then be used to accelerate the dynamics and/or explore the configurational space further. The advantage is that the Autoencoder is able to study systems without any prior chemical intuition, and searching for undiscovered states. In this work, we want to introduce these approaches and compare the two methods.

Recently, *Rizzi* et al.[23] used Deep-LDA to investigate the non-covalent interactions between a calixarene host and various guest molecules, selected as a model of more complex protein-ligand interactions. Most importantly, the role of water molecules was taken into account to identify the CVs and calculate the binding free energies. In our study, we first aim at applying both Deep-LDA and Autoencoder algorithms to the investigation of a complex biological system to give a comparison of the results obtained from both the methods. In general, the Autoencoder approach is more general, based on only basic coordinates of the system and does not require any prior knowledge of the system or the energy landscape. Nevertheless, it comes up with a set of CVs that accelerate the sampling of configuration space. Although, this approach is computationally more demanding than Deep-LDA, in a scenario where no prior knowledge is available, it can infer a faster and more accurate enhanced sampling run, as well as constraints, to sample the system more efficiently and with less predefined bias when selecting CVs.

Here, to test and compare the aforementioned methods, a well-researched (preclinical)-drug/target complex has been chosen, constituted by a pharmacologically relevant non-canonical DNA secondary structure (G-quadruplex) and a small molecule acting as its stabilizer. G-quadruplexes (G4 s) are secondary DNA structures formed in guanine rich sequences self-assembled by Hoogsteen-type hydrogen bonds and have been identified in human telomeres and promoter regions of many genes, where they regulate telomere homeostasis, gene transcription and DNA replication.[24] Stabilization of G4 s by small molecules has been shown to induce anticancer effects due to the resulting inhibition of telomere extensions or oncogene expression.[25]

In details, the organometallic gold complex ([Au(9-methylcaffein-8-ylidene)$_2$]$^+$) (**AuTMX$_2$**)[26] was selected as the stabilizer, and the oncogene promoter DNA G4 sequence cKIT1 (pdb 4WO2)[27] was the nucleic acid target. Previous metadynamics (metaD) studies have been applied by us on this ligand/G4 system and enabled the assessment of the binding modes and free-energy landscape of the **AuTMX$_2$**/cKIT1 non-covalent adduct.[28] Notably, the theoretical results were validated by experimental assays and provided an accurate estimate of the absolute gold compound/DNA binding free energy. The in silico results revealed two ligand binding modes for **AuTMX$_2$** on the G4 structure: one more thermodynamically favoured, which corresponds to **AuTMX$_2$** interaction with the top of the guanine tetrad, and the other one with the compound interacting with both the top of the tetrad and a flanking base.[28] These results were in agreement with previously reported X-ray diffraction (XRD) studies on the gold compound's adduct with another G4 model.[29]

Results and Discussion

Initially, we aimed at reproducing the previously obtained free energy values and binding modes with the help of the Deep-LDA-generated CV (s_w). Therefore, two states of the **AuTMX$_2$**/cKIT1 adduct were defined, namely bound (B, **AuTMX$_2$** interacting with the purine bases of the uppermost G4 tetrad) and unbound (U, **AuTMX$_2$** and cKIT1 not interacting at a distance beyond 2 nm) states.

The strategy for obtaining the binding free energy values, using Deep-LDA to determine the CVs, includes three main steps: i) generation of a number of descriptors, namely water coordination values for the solvation descriptors, to discriminate between B and U states (Figure 1A-a); ii) Deep-LDA deep neural network (DNN) in order to identify the optimal non linear discriminant function s_w of the descriptors above (Figure 1A-b)); iii) metaD calculations using s_w as abstract CV and the distance between K^+ and Au^+ ions ($d_{K\text{-}Au}$) as geometric CV, from which free energy surfaces (FES) are calculated (Figure 1A-c, see experimental for details). Following these steps, starting from the B and U states, 12 water solvation descriptors were defined: some of them cantered on the carbon atoms of the **AuTMX$_2$** caffeine ligands' and covering the area around them (L1–L4, Figure 1A-a), and others located in the area on top of the G4's uppermost tetrad (V1–V8, Figure 1A-a). This choice is based on our interest in capturing the role of water in the **AuTMX$_2$**/cKIT1 binding process. Starting from a full inclusion of 100% of all water atoms present (ca. 32000), Deep-LDA could find a projection of the latent space that was able to separate the states B and U satisfactorily (Figure S4).

To check its robustness and to assess simplifications of the method for a reduction of overall computational time, calculations were also performed with the inclusion of lower amounts of water, i.e. only ca. 40% (ca. 13200), 50% (ca. 16200) and 60% (ca. 19500) of all water molecules, respectively (see Figures S1–S4). To favour multiple binding events in short timescales, a funnel restraint was added on top of the G4 topmost tetrad, to retain **AuTMX$_2$** close to the DNA.[30] This funnel was constructed with an inner radius $r_{cyl}=2$ Å, opening up as a cone onto the cKIT1 tetrad (see Experimental for details). The obtained average free energy results are summarized in Figure 2 and the corresponding free energy surface (FES)

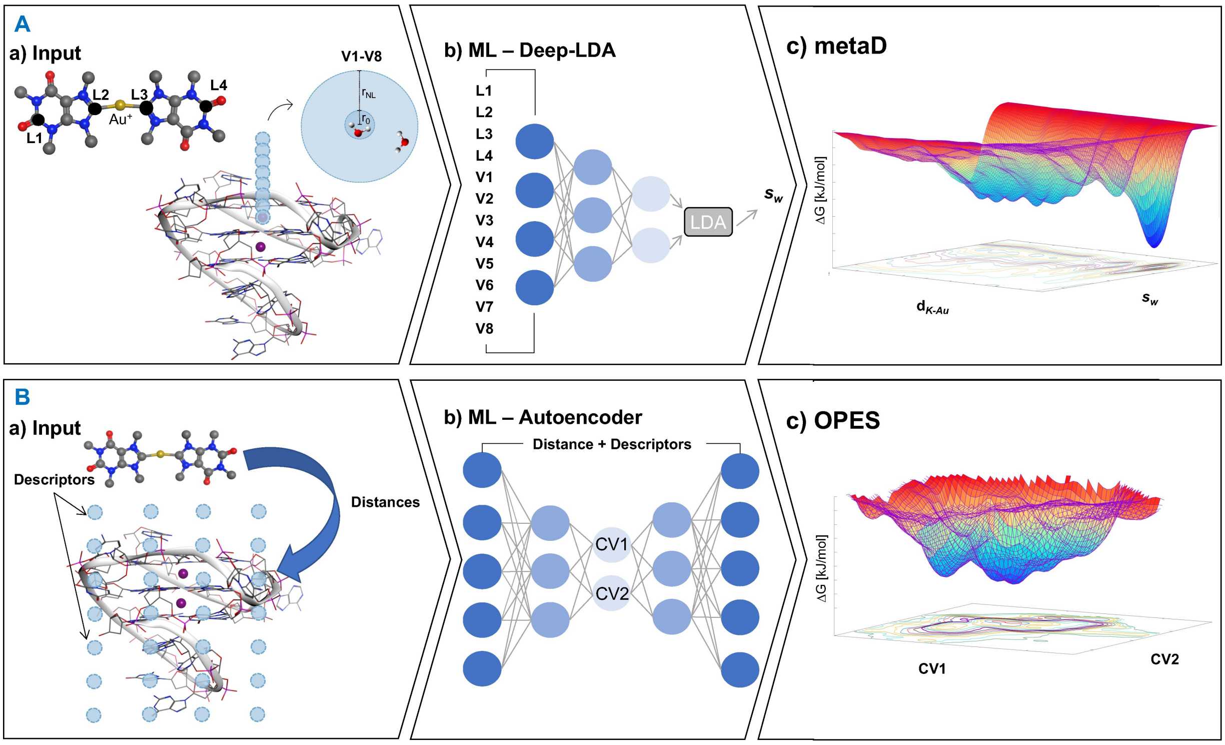

Figure 1. A) Schematic three-step process of the Deep-LDA based simulation: a) Input generation: Setup of virtual atoms on the Au(I) compound AuTMX_2 (L1–L4, black points), and on top of the upper tetrad of cKIT1 for water coordination (V1–V8, for better visibility, only the inner sphere radius r_0 is displayed for the arrangement of virtual atoms in the z-direction). b) Deep-LDA: Taking the descriptor values as input, Deep-LDA finds the projection of features that maximizes the separation between the bound (B) and unbound (U) states s_w. c) metaD: 3-dimensional free energy surface plot, generated from the final simulation using the Deep-LDA CV (s_w) and a distance CV ($d_{K\text{-}Au}$) as simulation bias. The global energy minimum (dark blue) corresponds to the bound state B. AuTMX_2 and water are shown in balls and sticks, the cKIT1 DNA is represented as sticks with ribbon. B) Schematic representation of Autoencoder method: a) Input generation: Inputs of the Autoencoder are all the distances between the heavy atoms of AuTMX_2 and the residues of cKIT1 and a grid of virtual atoms placed around the host molecule as solvation descriptors. b) Autoencoder network: The network is trained using unlabeled data from unbiased simulations with mean square error between the input and output as loss function c) OPES: On-the-fly probability enhanced sampling is performed using the latent space of the trained Autoencoder as CVs. A free energy surface of CVs can be calculated, and eventually, through the process of reweighting, binding energy of the AuTMX_2 interacting with cKIT1 can be approximated.

for the 100% water run is also shown. When all 100% water molecules were considered, our values show a slightly weaker, but very similar stabilization of the cKIT1 DNA (Figure 2A–B, ΔG = −37.0 ± 11.9 kJ/mol) in comparison to previous metaD reports (ΔG = −45 ± 3 kJ/mol).[28]

It should be noted that free energy values reported in Figure 2B were obtained by averaging the ΔG values of the most stable identified binding states in each run (see below). Further runs including 40%, 50% and 60% water during the calculations rendered binding free energy values of ΔG = −46.9 ± 2.3 kJ/mol, −41.6 ± 8.0 kJ/mol and -51.9 ± 2.7 kJ/mol, respectively (Figure 2B). All values are close to each other and in accordance to the previously performed metaD studies, proving the validity, as well as the robustness of the method and the resulting CV s_w.

The simulations originally obtained by metaD using the two empirical CVs rendered two main states of interaction between **AuTMX_2** and cKIT1.[28] The most stable state (**state I**) corresponded to π-π interactions between **AuTMX_2** and the guanine bases (G2, G6 and/or G10) of the topmost tetrad. The second state involved π-π interactions between **AuTMX_2** with guanines (G10 or G6) located on top of the upper tetrad, as well as on top of the flanking adenine base A1 (**state II**).[28]

As metaD explores the whole energy surface of an interaction, rather than just one minimum, multiple metastable positions can also be observed in addition to the global minimum. In the present study, similar binding states were found across the different simulation setups for our runs applying the CVs of our ML-based approach ((s_w) and the distance CV ($d_{K\text{-}Au}$)) (Figure 2 and Figures S4–S7). Specifically, the two states could be identified as the most recurring and stable conformations during our simulations, with a slight preference for **state II** over **state I** as most stable state. These results could also be reproduced consistently for runs with lower amounts of water (40% to 60%). Interestingly, and in contrast to previous findings,[28] **state II** was observed with a certain frequency as the one being marginally lower in energy (−42.6 ± 9.7 kJ/mol, averaged over all observed (metastable) states and setups, compared to −38.1 ± 12.4 kJ/mol for **state I**). This result can be explained as **state II** features a three-fold stack, based on π-π-interactions of **AuTMX_2** with A1, and of A1 with G2 and/or G13 located underneath. The (meta−)stable states for each simu-

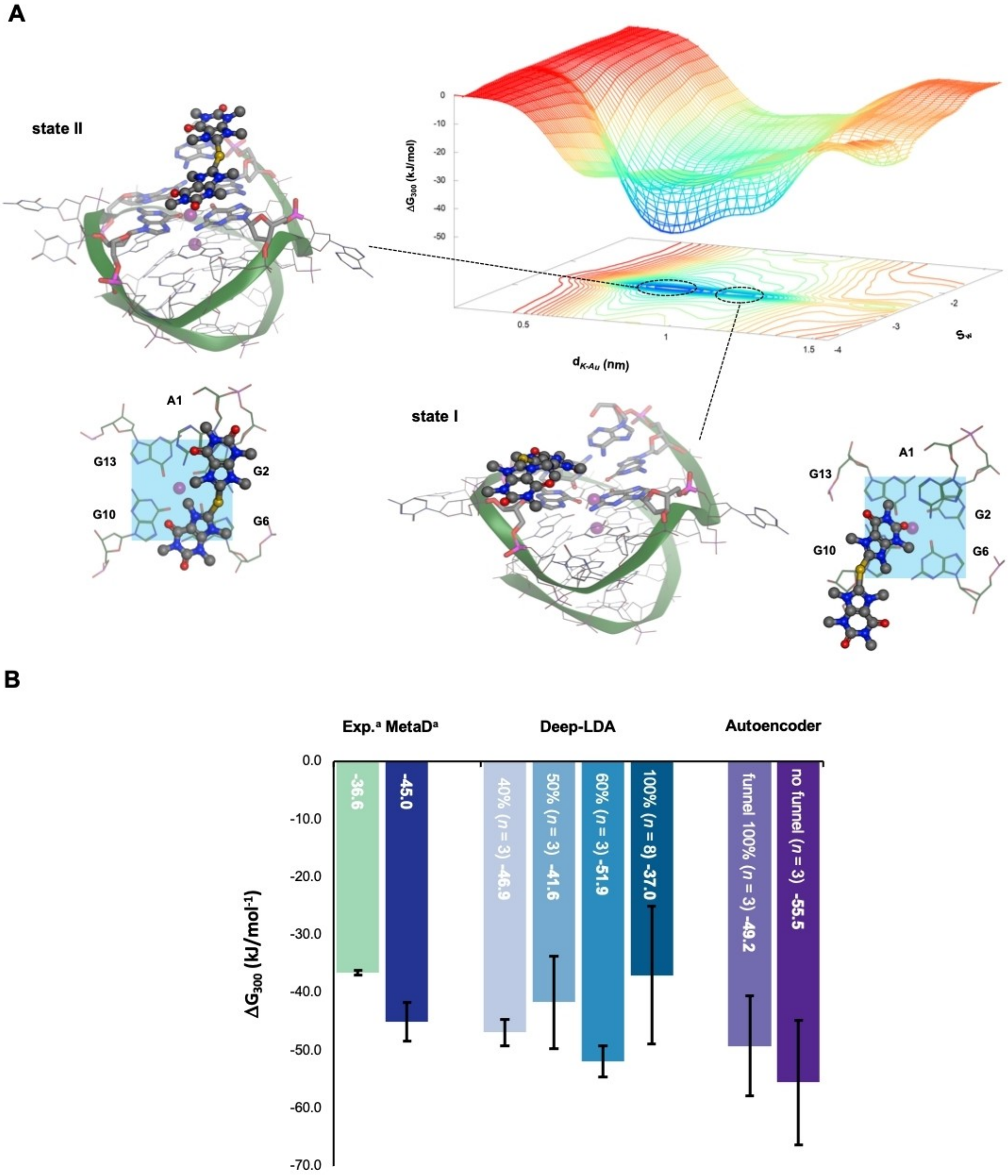

Figure 2. A) Free Energy Surface (FES) of the interactions between $AuTMX_2$ and cKIT1 calculated applying Deep-LDA-based metaD considering 100% water inclusion. CVs include the distance $d_{K\text{-}Au}$ CV (nm) and the ML-based water coordination CV s_w. The main binding poses (state I and II) of $AuTMX_2$ interacting with cKIT1 are shown. For each binding state, also the top view of the uppermost G4 tetrad is depicted with specific nucleobase residues highlighted. cKIT1 is shown in sticks representation with green backbone ribbon. $AuTMX_2$ is shown as balls and sticks. The chemical structures were created using the Molecular Operating Environment (MOE) software. B) Free energy values (ΔG) of the most stable states for the $AuTMX_2$/cKIT1 adduct obtained by different methods, including experimental data (DNA melting assay) or metaD approaches using empirically determined CVs, and ML-based approaches (Deep-LDA/Autoencoder), to obtain mean binding free energy values. *n* states the number of individual runs performed for calculating the average ΔG. The values are corrected for the presence of the funnel potential (SI Equation (2)). [a] Taken from reference [28].

lation setup can be found in the Supporting Information (Figures S4–S7 and Tables S1).

Afterwards, to assess the Autoencoder approach, training data based on an unbiased MD simulation of 4 ns, initialized starting from the unbound state of the ligand, were obtained. The sampling included from the beginning a solvation sphere around the DNA host molecule to include the water effect, effectively equivalent to include the solvent in the simulation. The input to the Autoencoder is a one-dimensional vector containing all the distances between the residues of the host and the heavy atoms of the ligand as well as a 3×3×4 grid of solvation descriptors uniformly distributed around the G4 molecule (Figure 1B-a). A total of 1236 descriptors are fed into a deep Autoencoder network and trained to minimize the mean square error between the input and the output. The system learns to distinguish the configurations along the distance between the cKIT1 structure and $\mathbf{AuTMX_2}$ (Figure S8). This is the first indication that the exploration of CV space will result in the exploration of bound and the unbound minima.

We have tested the Autoencoder in two scenarios: the first where a similar funnel potential as in the Deep-LDA simulation was added, and the second, where this constraint was removed. The first set of simulations was performed for the sake of comparison with the Deep-LDA results in order to test if the Autoencoder recovers similar binding sites and energies. The evaluation of the distance between **AuTMX$_2$** and cKIT1 during the simulations (Figure S9) show that the bound and the unbound states are visited multiple times which is crucial to get the correct statistics for the free energy calculations. Of note, the plots for the CVs (Figure S9) show that only a part of the whole CV space is explored during these simulations with the funnel potential, indicating that the same model has the capability to explore more configurations once the latter is removed.

The free energy surface with the funnel potential shows 4 different minima (Figure S10), where **AuTMX$_2$** interacts with the top tetrad of the cKIT1 in similar poses as those identified by the DeepLDA approach. The implication is that even if the CVs obtained by the Autoencoder are different in the two methods, the quality of the sampling is the same and they recover similar minima. In particular, the use of the same constraint in the two cases assures that both are sampling in the same portion of the configuration space. However, the Autoencoder started from more general coordinates showing a better generalization compared to Deep-LDA.

To assess the capability of the Autoencoder to identify correct CVs even when no information on the dynamic of the system is available, another simulation was conducted where the funnel potential was removed giving complete generality to the algorithm in sampling the configuration space (Figure S11). The resulting free energy surface from the 80 ns run is shown in Figure 3. The two free energy minima correspond to interactions of **AuTMX$_2$** with the top tetrad bases G6 and G13 ($\Delta G_{300}=-49.9$ KJ/mol) (**state I**) while the other minima actually show a different binding site at the bottom of cKIT1, where **AuTMX$_2$** is interacting with the G18 residue ($\Delta G_{300}=-43.4$ KJ/mol) (**state III**) (Figure 2B). **State I** is the one previously identified with the Deep-LDA and the simulation of the Autoencoder with the funnel. It represents a "collection" of very similar local minima which are not showed separately for clarity.

More interesting is the new **state III**, featuring a local minimum due to the π-interaction of the gold compound with the G18 residue. It represents the capability of the Autoencoder's CVs to explore unseen configurations starting from very general coordinates of the system. This new binding state can be observed by relaxing external constraints in the sampling region and CV selection. Obviously, this comes at the cost of a longer simulation to train the algorithm (80 ns versus 30 ns with the funnel).

Overall, the Autoencoder showed the same kind of π-π interaction between **AuTMX$_2$** and cKIT1 as detected by the Deep-LDA. However, another key binding site (state **III**) was also identified, which was not detectable even via classical XRD studies,[29] thus, corroborating the idea that data driven methods could favour the identification of alternative drug binding domains on even more elusive pharmacological targets. The insert on the right of Figure 3 shows a 3D map of the log

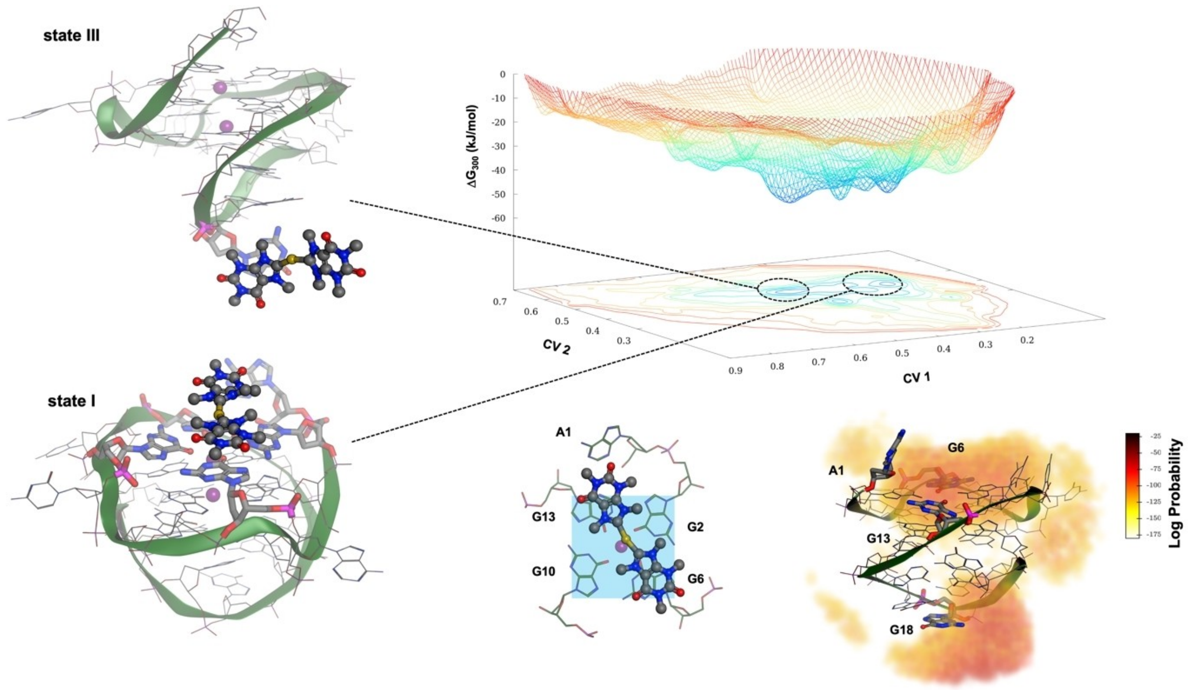

Figure 3. Free Energy Surface (FES) of interactions between AuTMX$_2$ and cKIT1 using Autoencoder CVs to accelerate the sampling. No funnel potential applied. For state I, also the top view of the uppermost G4 tetrad is shown with specific nucleobase residues highlighted. cKIT1 is shown in sticks representation with green backbone ribbon. AuTMX$_2$ is shown as balls and sticks. The chemical structures were created using the Molecular Operating Environment (MOE) software.

probabilities of discovering the **AuTMX$_2$** around cKIT1 in trajectories through the Autoencoder method, showing that only the input of general coordinates of the system results in calculation of a map of possible binding sites around the host. These potential binding sites can then be further studied by constraining the system in the proximity of a minimum and collecting more precise statistics through a secondary enhanced sampling run with funnel dynamics like in Deep-LDA.

Conclusions

In conclusion, we have reported here on an integrated ML approach to optimize the choice of the CVs for metaD enabling the investigation of a model (preclinical)-drug/target complex at a molecular level. The obtained results show that both of the selected algorithms – Deep-LDA and Autoencoder – have the capability to handle the large system and provide consistent results in terms of binding modes and free energies. The Autoencoder is more general but more expensive, due to the large training set. Deep-LDA is usually cheaper in this respect but requires more prior knowledge of the system. The Autoencoder showed the capability, starting from more basic coordinates (atomic distances, simple spatial grid to include the effect of the solvent), of deducing chemically relevant CVs. More importantly, it demonstrated the ability to test different constraints in order to have a better picture of possible minima in the host–guest system.

However, both methods rely on appropriate initialization of the initial time trajectories for sampling which is based on some chemical intuition. Thus, even if the Autoencoder can start from more general coordinates, still a preliminary decision concerning the training data has to be made. We also envision a hierarchical framework where an Autoencoder-based approach sits on top of Deep-LDA to study systems without any prior knowledge using only structural information. The information obtained from this can then be further used, if required, in Deep-LDA to perform informed exploration of transition states.

A third level in the hierarchy would be one in which a data-driven method samples the training data assuming some "thermodynamic constraint" i.e., the sampled data reproduce a Maxwell-Boltzmann distribution. Similar methods are under investigation[31] and can easily be integrated within our hierarchical concept to make the CV selection more and more unbiased.

Overall, our work paves the way to key applications of ML in drug discovery enriching the toolbox of methods available for computer-aided drug design (CADD), beyond quantitative structure-activity relationship (QSAR) analysis, virtual screening and de novo drug design.[32] Specifically, data-driven methods will boost the unbiased elucidation of target-ligand interactions at an atomic level, thus, improving the drug design strategy. Furthermore, our approach enables the inclusion of water molecules in the binding process, affecting also the target structure and ligand solvation. In the future, this data-driven approach can overcome most of the current limitations in the use of atomistic simulations to study noncovalent interactions beyond drug/target interactions; for example, in the characterization of the host guest-chemistry of supramolecular materials in different environments.

Supporting Information

The authors have cited additional references within the Supporting Information.[33–45]

Acknowledgements

The authors acknowledge support from the TUM Innovation Network "Artificial Intelligence Powered Multifunctional Material Design" (ARTEMIS). Open Access funding enabled and organized by Projekt DEAL.

Conflict of Interests

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords: collective variables · machine learning · metadynamics · metallodrugs · G-quadruplexes

[1] W. L. Jorgensen, *Science* **2004**, *303*, 1813–1818.

[2] M. De Vivo, M. Masetti, G. Bottegoni, A. Cavalli, *J. Med. Chem.* **2016**, *59*, 4035–4061.

[3] B. Peters, *Reaction Rate Theory and Rare Events*, Elsevier, Oxford, **2017**, p. 3-5.

[4] K. Lindorff-Larsen, S. Piana, R. O. Dror, D. E. Shaw, *Science* **2011**, *334*, 517–520.

[5] Y. I. Yang, Q. Shao, J. Zhang, L. Yang, Y. Q. Gao, *J. Chem. Phys.* **2019**, *151*, 070902.

[6] a) W. L. Jorgensen, L. L. Thomas, *J. Chem. Theory Comput.* **2008**, *4*, 869–876; b) W. L. Jorgensen, C. Ravimohan, *J. Chem. Phys.* **1985**, *83*, 3050–3054.

[7] G. M. Torrie, J. P. Valleau, *J. Comput. Phys.* **1977**, *23*, 187–199.

[8] Y. Sugita, Y. Okamoto, *Chem. Phys. Lett.* **1999**, *314*, 141–151.

[9] A. Laio, M. Parrinello, *Proc. Nat. Acad. Sci.* **2002**, *99*, 12562–12566.

[10] a) H. Grubmüller, B. Heymann, P. Tavan, *Science* **1996**, *271*, 997–999; b) B. Isralewitz, M. Gao, K. Schulten, *Curr. Opin. Struct. Biol.* **2001**, *11*, 224–230.

[11] D. Hamelberg, J. Mongan, J. A. McCammon, *J. Chem. Phys.* **2004**, *120*, 11919–11929.

[12] A. K. Faradjian, R. Elber, *J. Chem. Phys.* **2004**, *120*, 10880–10889.

[13] P. G. Bolhuis, D. Chandler, C. Dellago, P. L. Geissler, *Annu. Rev. Phys. Chem.* **2002**, *53*, 291–318.

[14] G. Bussi, A. Laio, *Nat. Rev. Phys.* **2020**, *2*, 200–212.

[15] F. L. Gervasio, A. Laio, M. Parrinello, *J. Am. Chem. Soc.* **2005**, *127*, 2600–2607.

[16] a) D. Wragg, A. de Almeida, A. Casini, S. Leoni, *Chem. Eur. J.* **2019**, *25*, 8713–8718; b) D. Wragg, S. Leoni, A. Casini, *RSC Chem. Biol.* **2020**, *1*, 390–394; c) A. Cavalli, A. Spitaleri, G. Saladino, F. L. Gervasio, *Acc. Chem. Res.* **2015**, *48*, 277–285; d) F. S. Di Leva, E. Novellino, A. Cavalli, M. Parrinello, V. Limongelli, *Nucleic Acids Res.* **2014**, *42*, 5447–5455; e) A. D. Favia, M. Masetti, M. Recanatini, A. Cavalli, *PLoS One* **2011**, *6*, e25375;

f) M. Masetti, A. Cavalli, M. Recanatini, F. L. Gervasio, *J. Phys. Chem. B* **2009**, *113*, 4807–4816.

[17] a) J. Debnath, M. Parrinello, *J. Phys. Chem. Lett.* **2020**, *11*, 5076–5080; b) L. Bonati, Y.-Y. Zhang, M. Parrinello, *Proc. Nat. Acad. Sci.* **2019**, *116*, 17641–17647; c) M. Invernizzi, M. Parrinello, *J. Phys. Chem. Lett.* **2020**, *11*, 2731–2736; d) O. Valsson, M. Parrinello, *Phys. Rev. Lett.* **2014**, *113*, 090601.

[18] a) L. Bonati, V. Rizzi, M. Parrinello, *J. Phys. Chem. Lett.* **2020**, *11*, 2998–3004; b) A. Barducci, M. Bonomi, M. Parrinello, *WIREs Comput. Mol. Sci.* **2011**, *1*, 826–843.

[19] a) F. Noé, C. Clementi, *Curr. Opin. Struct. Biol.* **2017**, *43*, 141–147; b) D. Mendels, G. Piccini, M. Parrinello, *J. Phys. Chem. Lett.* **2018**, *9*, 2776–2781; c) G. Piccini, D. Mendels, M. Parrinello, *J. Chem. Theory Comput.* **2018**, *14*, 5040–5044.

[20] a) M. M. Sultan, V. S. Pande, *J. Chem. Phys.* **2018**, *149*, 094106; b) Z. Belkacemi, P. Gkeka, T. Lelièvre, G. Stoltz, *J. Chem. Theory and Computation* **2022**, *18*, 59–78; c) B. Luigi, P. GiovanniMaria, P. Michele, *Proc. Nat. Acad. Sci.* **2021**, *118*, e2113533118; d) S. Brandt, F. Sittel, M. Ernst, G. Stock, *J. Phys. Chem. Lett.* **2018**, *9*, 2144–2150; e) W. Chen, A. L. Ferguson, *J. Comput. Chem.* **2018**, *39*, 2079–2102; f) W. Chen, H. Sidky, A. L. Ferguson, *J. Chem. Phys.* **2019**, *150*, 214114; g) C. Wehmeyer, F. Noé, *Chem. Phys.* **2018**, *148*, 241703.

[21] Y. Bengio, A. Courville, P. Vincent, *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828.

[22] W. Chen, A. R. Tan, A. L. Ferguson, *J. Chem. Phys.* **2018**, *149*, 072312.

[23] V. Rizzi, L. Bonati, N. Ansari, M. Parrinello, *Nat. Commun.* **2021**, *12*, 93.

[24] D. Varshney, J. Spiegel, K. Zyner, D. Tannahill, S. Balasubramanian, *Nat. Rev. Mol. Cell Biol.* **2020**, *21*, 459–474.

[25] S. Neidle, *Nat. Chem. Rev.* **2017**, *1*, 0041.

[26] S. M. Meier-Menches, B. Neuditschko, K. Zappe, M. Schaier, M. C. Gerner, K. G. Schmetterer, G. Del Favero, R. Bonsignore, M. Cichna-Markl, G. Koellensperger, A. Casini, C. Gerner, *Chem. Eur. J.* **2020**, *26*, 15528–15537.

[27] D. Wei, J. Husby, S. Neidle, *Nucleic Acids Res.* **2015**, *43*, 629–644.

[28] D. Wragg, A. de Almeida, R. Bonsignore, F. E. Kuhn, S. Leoni, A. Casini, *Angew. Chem. Int. Ed. Engl.* **2018**, *57*, 14524–14528.

[29] C. Bazzicalupi, M. Ferraroni, F. Papi, L. Massai, B. Bertrand, L. Messori, P. Gratteri, A. Casini, *Angew. Chem. Int. Ed.* **2016**, *55*, 4256–4259.

[30] F. Moraca, J. Amato, F. Ortuso, A. Artese, B. Pagano, E. Novellino, S. Alcaro, M. Parrinello, V. Limongelli, *Proc. Nat. Acad. Sci.* **2017**, *114*, E2136–E2145.

[31] M. Schöberl, N. Zabaras, P.-S. Koutsourelakis, **2020**, arXiv preprint DOI: 10.48550/arXiv.2002.10148.

[32] S. Dara, S. Dhamercherla, S. S. Jadav, C. M. Babu, M. J. Ahsan, *Artif. Intell. Rev.* **2022**, *55*, 1947–1999.

[33] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, D. E. Shaw, *Proteins* **2010**, *78*, 1950–1958.

[34] S. Nosé, *Mol. Phys.* **2006**, *52*, 255–268.

[35] M. Parrinello, A. Rahman, *Phys. Rev. Lett.* **1980**, *45*, 1196–1199.

[36] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, E. Lindahl, *SoftwareX* **2015**, *1–2*, 19–25.

[37] G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, G. Bussi, *Comput. Phys. Commun.* **2014**, *185*, 604–613.

[38] V. Limongelli, M. Bonomi, M. Parrinello, *Proc. Nat. Acad. Sci.* **2013**, *110*, 6358–6363.

[39] P. Xanthopoulos, P. M. Pardalos, T. B. Trafalis in *Robust Data Mining*, Springer, New York, **2013**, pp. 27–33.

[40] M. K. Dorfer, R. Kelz, G. Widmer, *ICLR* **2015** eprint arXiv:1511.04707..

[41] S. Bhakat, *RSC Adv.* **2022**, *12*, 25010–25024.

[42] a) S. Rifai, P. Vincent, X. Muller, X. Glorot, Y. Bengio, in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, Omnipress, Bellevue, Washington, USA, **2011**, pp. 833–840; b) P. Vincent, H. Larochelle, Y. Bengio, P.-A. Manzagol, in *Proceedings of the 25th international conference on Machine learning*, Association for Computing Machinery, Helsinki, Finland, **2008**, pp. 1096–1103; c) H. Lee, C. Ekanadham, A. Y. Ng, in *Proceedings of the 20th International Conference on Neural Information Processing Systems*, Curran Associates Inc., Vancouver, British Columbia, Canada, **2007**, pp. 873–880.

[43] a) J. Baima, A. M. Goryaeva, T. D. Swinburne, J.-B. Maillet, M. Nastar, M.-C. Marinica, *Phys. Chem. Chem. Phys.* **2022**, *24*, 23152–23163; b) W. Chen, A. L. Ferguson, *J. Comput. Chem.* **2018**, *39*, 2079–2102; c) W. Chen, A. R. Tan, A. L. Ferguson, *J. Chem. Phys.* **2018**, *149*, 072312; d) M. M. Sultan, V. S. Pande, *J. Chem. Phys.* **2018**, *149*, 094106.

[44] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, 1 ed., Routledge, **1998**.

[45] T. W. Allen, O. S. Andersen, B. Roux, *Proc. Nat. Acad. Sci.* **2004**, *101*, 117–122.

Manuscript received: July 28, 2023
Accepted manuscript online: August 9, 2023
Version of record online: September 28, 2023