QL-based Adaptive Transceivers for the IoBNT Communications

Roya Khanzadeh[1], Stefan Angerbauer[2], Jorge Torres Gomez[3], Andreas Springer[2], Falko Dressler[3], and Werner Haselmayr[2]

[1]Johannes Kepler University Linz, Institute for Communications Engineering and RF-Systems, JKU LIT SAL eSPML Lab, Linz, Austria
[2]Johannes Kepler University Linz, Institute for Communications Engineering and RF-Systems, Linz, Austria
[3]TU Berlin, School of Electrical Engineering and Computer Science, Berlin, Germany

***Abstract*—This paper introduces an adaptive transceiver scheme for bio-nano things (NTs) situated within blood vessels communicating through a time-varying molecular channel. The proposed scheme employs a Q-learning-based adaptive transceiver (a so-called QL-ADT), wherein an agent gradually learns how to adapt the transmission parameters to the current state of the channel. A real heart rate dataset is used to estimate the blood flow velocities over time, based on which a time-varying molecular channel is modelled. In the practical implementation of the QL-ADT, an external gateway, situated on the skin, monitors the body's heart rate over time and interfaces with the NTs through implantable nano devices. The gateway dynamically adjusts the communication parameters of the NTs based on the measured heart rate and what it has learned during the training phase. The proposed QL-ADT scheme showed significant improvement in the achievable raw bit rate (RBR) and error performance for a real heart rate dataset.**

***Index Terms*—Internet of Bio-Nano Things, Reinforcement Learning, Molecular Communications, Adaptive Transceivers.**

I. Introduction

THE internet of bio-nano things (IoBNT) is envisioned as a heterogeneous communication network, extending connectivity and control to unconventional domains like the human body with innovative applications in smart drug delivery and continuous health monitoring [1], [2]. There are mainly three types of devices involved in the IoBNT networks, namely biological or synthetic nano-things (NT), implantable nano-devices (IND) and nano-micro interfaces or gateways (GW). The NTs are resource-limited sensors and actuators located inside the human body communicating inside the human circulatory system (HCS) using molecular communication (MC). INDs are also located inside the human body and not only communicate with the NTs using MC, but they can also communicate with GWs through conventional electromagnetic waves (EM) such as THz signals. The more advanced devices in the IoBNT network are gateways, located outside the body but usually attached to the skin. Gateways are almost boundless-resource devices communicating through EM signals and transferring data inside and outside the human body [3].

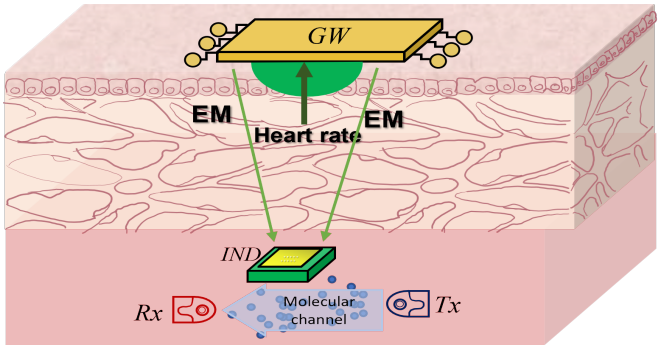

Fig. 1: The proposed system model for adaptive transceivers.

Concentration-based MC is commonly employed for communication between the NTs within the HCS, where information is encoded in the concentration levels of molecules which propagate through the blood vessels. The dispersion and transport of signalling molecules through the blood vessels are affected by time-varying blood flow velocity in the HCS, influencing not only molecules' concentration profiles (by making a shift in the peak of the concentration profile, called advection effect) but also the likelihood of overlap between successive symbols resulting in inter-symbol interference (ISI), called diffusion effect. In practice, the blood flow velocity within the HCS varies over time, particularly with changes in heart rate due to physical activities [4]. Communication performance degradation is the direct result of time-varying changes in the blood vessels as communication channels, and adaptive transmission is a promising solution for maintaining a required performance in time-varying channels.

Adaptive receiver (Rx) techniques, such as adaptive thresholding, have been vastly investigated for MC, like in [5], [6]. An adaptive transmission rate for binary signalling is also proposed in [7], where the knowledge of the prior transmitted symbols is used at the transmitter (Tx) to adjust the current symbol duration time. In these studies, the computation complexity of adaptive schemes is imposed on resource-constrained NTs, overlooking the potential for offloading this complexity to more sophisticated devices involved in the

network, like GWs. Therefore, this paper addresses a critical question: How can we leverage the advantages of adaptive transmission techniques without overloading resource-limited NTs with additional computational complexity, and instead offloading the complexity to an external GW?

We introduce an adaptive transceiver scheme for MC between NTs through a time-varying molecular channel caused by variations in blood flow velocities. The proposed scheme leverages the Q-learning (QL) algorithm, which is a particular reinforcement learning (RL) technique [8], wherein an agent stored in a GW interacts with the environment and learns through trial and error how to adjusts the transmission parameters to the fluctuating conditions of the molecular channel. To facilitate the practical implementation of the proposed approach, we have divided it into two distinct phases, namely the offline phase and the online phase.

In the offline phase, an agent learns how to select the transmission parameters for different molecular channel conditions by maximizing the rewards it gets for the selected parameters. This reward is designed to be proportional to the obtained performance of the current transmission in terms of error and data rate. Heart rate is the parameter that shows the channel condition for each transmission block. An electric-circuit model (referred to as a digital twin (DT)) replica of the HCS is used, enabling real-time estimation of corresponding blood flow velocity for each heart rate value, resulting in estimation of the current channel impulse response (CIR). Once the CIR is available, the transmission can be simulated on a computer and the performance parameters are evaluated and the corresponding reward is calculated. The agent gradually learns how to select the optimal transmission parameters (i.e. modulation order and symbol duration time) at each heart rate (or equivalently each CIR) to achieve the maximum reward (or equivalently maximum raw bit rate (RBR) while keeping the symbol error ratio (SER) within an acceptable range). The output of the offline phase is a Q-table showing the learned optimum transmission parameters for each heart-rate value. In the online phase, an external GW measures the current heart rate based on which the agent selects optimal transmission parameters using the trained model from the offline phase. The selected parameters for the current transmission are then fed back from the GW to the IND EM signals [9] and from the IND to the NTs using MC (see Fig. 1). Since both Rx and Tx dynamically adjust their parameters to the current transmission, we refer to the proposed scheme as an adaptive transceiver.

The main contributions of this study are summarized as follows:

- We propose a QL-based adaptive transmission for time-varying molecular communication.
- The proposed adaptive transmission offloads computational complexity to an external gateway, enhancing feasibility by alleviating the computational load on resource-limited nano-devices within the human body.
- We use a real heart rate dataset to validate the performance of the proposed scheme.

II. System Model

A concentration-based MC between a pair of NTs located within an artery is considered. Using the EM signals and through an IND, a GW outside the human body, which is attached to the human skin, transmits signals to these NTs. This GW has a sensor to monitor the heart rate in real-time, based on which an agent stored in the GW can be trained to select optimal parameters for current transmission between the NTs to maximize the possible transmission rate while keeping SER in an acceptable range. Fig. 1 illustrates the proposed system model.

Having access to the heart rate, a DT replica of the HCS is trained using the method in [4]. The obtained DT model then is used to estimate the blood flow velocities in real-time. At each transmission block of τ, the GW measures the heart rate and passes it to the DT model to estimate the corresponding blood flow velocity $v(\tau)[\mathrm{ms}^{-1}]$ at a specific position in the artery. It is assumed that during each transmission block, the heart rate does not change significantly. Considering that the two NTs are placed at a fixed distance d from each other, the corresponding time-varying CIR at the τth transmission block is given by

$$h(t,\tau) = \frac{1}{\sqrt{4\pi D_{\text{eff}} t}} e^{\frac{-(d-v(\tau)t)^2}{4D_{\text{eff}}t}}, \tag{1}$$

with $D_{\text{eff}} = D + \frac{r^2 v(\tau)^2}{48D}$, with the diffusion coefficient [10] $D = 2.75 \times 10^{-9}\,\mathrm{m^2 s^{-1}}$, the distance $d = 10^{-1}\,\mathrm{m}$ and the artery diameter [11] $r = 10^{-4}\,\mathrm{m}$. It is also assumed that the Tx uses adaptive M-ary concentration shift keying (CSK) modulation with adjustable symbol duration time. The molecules then pass through the molecular channel with the described CIR and are received at the Rx. To keep the computational complexities at the NTs as low as possible, a simple threshold detection is used at the Rx.

It is worth mentioning that the reason why modulation order and symbol duration time are chosen for adaptation is that both parameters contribute to communication performance in molecular channels with diffusion and advection characteristics. It means they directly influence the ability to differentiate symbols and mitigate the effects of overlapping signals passing through a channel with memory (diffusion effect) and varying peak arrival time of the received signal (advection effect). Hence, we expect to achieve higher transmission rates under conditions where the spread and delay of the signal in the channel are minimal, or equivalently when the heart rate is higher, as will be shown in Section IV.

III. The proposed Reinforcement Learning Based Adaptive Transceiver

The proposed method is based on the RL technique, wherein an agent learns optimal behaviour through interaction with its environment. In this letter, we utilize an off-policy temporal

difference algorithm called the QL algorithm. The QL is a model-free algorithm, meaning that it does not rely on an explicit model of the environment. This characteristic makes it a particularly suitable choice for addressing adaptive transmission problems [12].

At each training step n of the QL algorithm, the agent observes the state $s_n \in \mathcal{S}$ of the environment and selects an action $a_n \in \mathcal{A}$. As a consequence of this action, the agent receives a reward $r_n \in \mathcal{R}$ and perceives a new state s_{n+1}. The primary objective of the RL agent is to identify actions that yield a high reward. The agent determines the best actions for each observation by considering the value of an action-value function, denoted as $Q(s_n, a_n)$, or Q-function. This function represents the cumulative expected reward for taking an action in state s_n and subsequently following a policy. At each step, the estimation of the action-value function is updated by

$$Q(s_n, a_n) \leftarrow (1-\alpha)Q(s_n, a_n) + \alpha\big[r_n + \gamma \max_{a_{n+1}\in\mathcal{A}} Q(s_{n+1}, a_{n+1})\big], \tag{2}$$

where $0 \leq \alpha \leq 1$ and $0 \leq \gamma \leq 1$ are learning rate and discount factor respectively which should be designed correctly for each optimization task [8]. Therefore, a Q-table where the Q-values for each state-action pair are represented is also updated iteratively.

A. State Space

The state space in our problem consists of the heart rate values measured at the GW. Considering integer values for the heart rate, the state space, therefore, is discrete. The state at the training step n is denoted as s_n, which is equal to the corresponding heart rate for the current transmission interval.

B. Action Space

The action space consists of all the transmission parameters that can be adapted at each transmission interval and is denoted by

$$\mathcal{A} = \{(M_1, T_1), (M_2, T_1), \ldots, (M_K, T_J)\}, \tag{3}$$

where $M_k, k = 1, \ldots, K$ are the different available CSK modulation orders and $T_j, j = 1, \ldots, J$ are available symbol times. At step n, the taken action is $a_n = (M^n, T^n)$ with $M^n \in M_1, \ldots M_K$ and $T^n \in T_1, \ldots T_J$.

C. Reward

The reward function considered in this study is a non-linear function of the achievable RBR and error performance of the communication. The highest RBR, i.e. the highest modulation order and lowest symbol duration time, gets the highest reward, but only if the communication error is acceptable by the application. If the error performance of the chosen action is not acceptable, the reward would be negative. The error performance is measured based on the SER, and then the reward function at step n is defined as:

$$r_n = \begin{cases} 10\ \frac{\log_2 M^n}{T^n}(1 - \mathrm{SER}_n) & \text{if } \mathrm{SER}_n < \mathrm{thr} \\ -10\ \mathrm{SER}_n & \text{else.} \end{cases} \tag{4}$$

where thr is the acceptable threshold for error performance at the Rx. Thus, with this reward function, the agent will try to maximize the RBR while ensuring error performance remains below a threshold and minimizing errors wherever feasible.

D. QL-based Adaptive Transmission Algorithm (QL-ADT)

The proposed approach is referred to as QL-based adaptive transmission or QL-ADT. The QL-ADT approach is divided into two phases: the offline phase and the online phase.

1) Offline Phase: In the offline phase, the agent is expected to explore the environment and learn how to obtain the greatest possible reward at each state (for each heart rate). In other words, in the offline phase, we train the agent iteratively, using a heart rate dataset. At each iteration, the agent gets a sample from the heart rate dataset as the current state s_n, and takes a step by selecting a proper action a_n, i.e. modulation order and symbol time, using its policy, and then updates its Q-table based on the obtained reward. Each step corresponds to an action chosen from the action space following a policy. An ϵ-greedy policy is used for this task which balances exploration and exploitation by selecting the best action with high probability and exploring with a smaller probability ϵ [13]. The approach fine-tunes the balance between exploring new possibilities and exploiting learned knowledge, enhancing decision-making in uncertain environments. The parameter ϵ will be initialized with ϵ_{max} which is some value close to one and then decays over time until it reaches ϵ_{min} which is a value close to zero.

Each iteration terminates if either the agent takes n_{max} steps or it receives the maximum expected reward r_{max}. By defining n_{max}, we encourage the agent to explore the environment more and take actions which might lead to higher rewards. On the other hand, r_{max} prevents the agent from becoming overwhelmed by excessive searching, which reduces the complexity of the training phase. The final output of the offline phase is a Q-table, showing the value of the actions for each state. A pseudo-code of the offline phase of the proposed QL-ADT algorithm is provided in Algorithm 1.

2) Online Phase: In the online phase, a GW attached to the human skin measures the heart rate for the current transmission block as s_τ and selects a proper action using the trained Q-table during the offline phase. These selected actions, i.e. $a^*_\tau = (M^\tau, T^\tau)$, then are transmitted from the GW to the NTs (both Tx and Rx) using the EM communications. The Tx and Rx then adapt their parameters for the current transmission block.

IV. Simulation Results

A. Digital Twin Model

For the DT model, we use the electric circuit representation of the HCS, as given in our previous work in [4]. This model

Algorithm 1 QL-ADT Algorithm

Initialization $Q(s_n, a_n) \leftarrow 0$ for all s_n and a_n
Input Heart-rate dataset
Output Optimal Q-table
Loop for each iteration:
- Agent gets a sample from the heart-rate dataset as s.
- $n \leftarrow 1$.
While $r_n \leq r_{\max}$ and $n \leq n_{\max}$:
- For $s_n = s$, the GW chooses a_n from $\mathcal{A}$ based on ϵ-greedy policy and the available Q-table.
- Tx transmits the signal using a_n.
- Rx detects the received signal using a_n and calculates SER_n.
- Agent calculates the reward and updates the Q-table using Eq. (4) and Eq. (2).
- $n \leftarrow n + 1$.
- go to the next iteration.

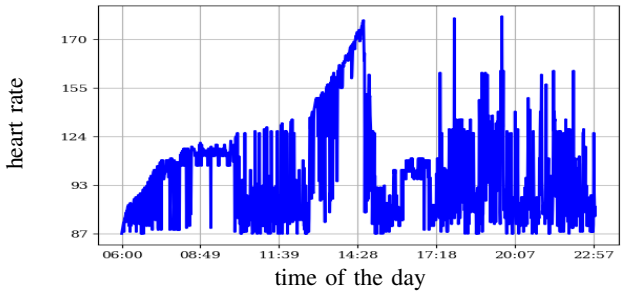

Fig. 2: The measured heart rate over a day, from 6 to 23.

allows us to evaluate the pressure as a function of time at different vessel segments for an arbitrary heartbeat frequency. Specifically, with this model, we evaluate the pressure in the wrists and use [4, Eq. (7)] to later evaluate the blood flow velocity at an arbitrary vessel stream. In other words, the input of the DT model is the heart rate value and its output is the corresponding estimated blood flow velocity in the considered artery in the wrist, with diameter mentioned in Section II. Based on the obtained velocity, we can estimate the corresponding CIR. It should be mentioned that estimating the blood flow velocity for each single heart rate value using the DT model is time-consuming. Therefore, we utilized the DT model to obtain flow velocities for selected values of the heart rate and then estimated the missing values using linear interpolation.

B. Simulation Parameters

Table I presents an overview of the main simulation parameters. It is assumed that the transceiver supports four modulation orders as 2-CSK, 4-CSK, 8-CSK and 16-CSK, along with five symbol duration time as 1s, 0.9s, 0.8s, 0.7s, and 0.6s [1]. Figure 3 illustrates examples of estimated blood flow velocities

[1]The symbol durations are strategically selected in accordance with the CIRs illustrated in Fig. 3

TABLE I: Simulation parameters

Parameter	Value
$M_k, k = 1, \ldots, 4$	2-CSK, 4-CSK, 8-CSK, 16-CSK
$T_j, j = 1, \ldots, 5$	1 s, 0.9 s, 0.8 s, 0.7 s, 0.6 s
learning rate α	0.05
discount factor γ	0.1
SER threshold (thr)	5×10^{-2}
maximum exploration rate ($\epsilon_{\max}$)	1
minimum exploration rate ($\epsilon_{\min}$)	0.001
$n_{\max}$	10
$r_{\max}$	40
Iterations	5×10^5

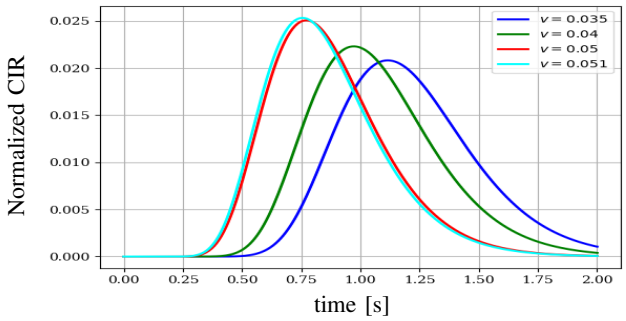

Fig. 3: The estimated CIRs for four different heart rate values, where $v = 0.035$, 0.04, 0.05 and 0.051 respectively corresponds to the heart rate values 87, 100, 160 and 180.

from the DT model, i.e. v, alongside their corresponding CIRs for four different heart rate values, which are normalized in power. The figure clearly illustrates the impact of varying flow velocities on both diffusion and advection processes, as evidenced by the extended duration of the CIRs and the noticeable shift in peak arrival times with different velocities. A real heart rate data set measured from a test person who wore a Garmin Fenix x5 smartwatch on his left wrist over a day from 6 to 23 is used for evaluation. The corresponding data is depicted in Fig. 2.

C. Results Analysis

Fig. 4 shows the total reward the agent earns at each iteration during the training. At the beginning of the training process, the received reward varies significantly, because the agent explores more and takes more random actions in the current state to avoid getting stuck in local optima. However, the agent gradually learns the association between states and actions after a sufficient number of iterations.

To assess the performance of the proposed QL-ADT algorithm, we evaluated the obtained SER and RBR (in bits per second [bps]) for the heart rate dataset illustrated in Fig. 2. In Fig. 5, the instantaneous achieved RBR by QL-ADT is compared with that of various fixed transmission schemes. As we expected, it is evident from this figure that the higher transmission rates are achieved by the proposed algorithm where the heart rate rises. Specifically, at around time $14:28$, when there is a significant peak in the heart rate, the QL-ADT algorithm shows a sharp peak in the achieved RBR. It should

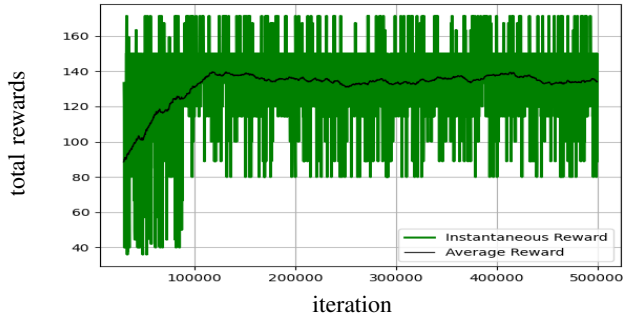

Fig. 4: Total Reward per iteration in QL-ADT algorithm.

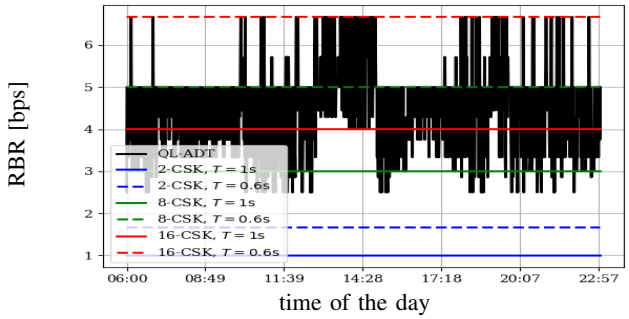

Fig. 5: The achieved RBR by QL-ADT for the heart rates in Fig. 2 compared to fixed-transmission schemes.

be noticed that adjusting symbol duration downward and/or boosting modulation order can both elevate the achieved RBR; however, escalating the modulation order again intensifies interference. Thus, the agent is expected to find the optimal balance for each observation, i.e. heart rate value.

We compared the average performance of the proposed QL-ADT method with fixed transmission schemes and summarized them in Table II. It is evidence from the table that QL-ADT archives higher average RBR compared to 2-CSK, (8-CSK, $T = 1\,\mathrm{s}$) and (16-CSK, $T = 1\,\mathrm{s}$), but still remains the required error performance (which is set to 5×10^{-2} Table I). In addition, Table II reveals although (8-CSK, $T = 0.6\,\mathrm{s}$ and (16-CSK, $T = 0.6\,\mathrm{s}$ can potentially convey higher data rate, they cannot maintain the required SER for the entire transmission (due to the intensified ISI effect in low-heart rate regimes).

These results confirm that using fixed transmission parameters to communicate through a time-varying MC channel leads to a compromise between a low bit rate and a high error ratio. In contrast, the proposed QL-ADT effectively identifies the optimal balance between bit rate and error ratio tailored to each specific channel condition, thereby facilitating more reliable communication.

V. Conclusion

In this letter, we proposed an adaptive transceiver design for MC based on QL algorithm. An external agent monitors heart rate and gradually learns to optimize transmission parameters for efficient MC between NTs. Our simulations demonstrate

TABLE II: The average achievable SER and RBR by QL-ADT compared to fixed-transmission schemes.

Scheme	**avg RBR [bps]**	**avg SER**
QL-ADT	4.54	2×10^{-2}
2-CSK, $T = 1\,\mathrm{s}$	1	0
2-CSK, $T = 0.6\,\mathrm{s}$	1.66	0
8-CSK, $T = 1\,\mathrm{s}$	3	0
8-CSK, $T = 0.6\,\mathrm{s}$	5	0.68
16-CSK, $T = 1\,\mathrm{s}$	4	0.27
16-CSK, $T = 0.6\,\mathrm{s}$	6.66	0.83

superior row bit rate and error performance when compared to fixed transmission schemes. Future work could explore extensions to incorporate the impact of NTs mobility within the environment.

Acknowledgment

This work has been in part supported by the "University SAL Labs" initiative of Silicon Austria Labs (SAL) and its Austrian partner universities for applied fundamental research for electronic-based systems as well as by the project IoBNT, funded by the Federal Ministry of Education and Research (BMBF, Germany) under grant 16KIS1986K.

References

[1] I. Akyildiz *et al.*, "The Internet of Bio-Nano Things," *IEEE Communications Magazine*, vol. 53, no. 3, pp. 32–40, Mar. 2015.

[2] F. Dressler and S. Fischer, "Connecting In-Body Nano Communication with Body Area Networks: Challenges and Opportunities of the Internet of Nano Things," *Elsevier Nano Communication Networks*, vol. 6, pp. 29–38, 6 2015.

[3] K. Yang, D. Bi, Y. Deng, R. Zhang, M. M. U. Rahman, N. A. Ali, M. A. Imran, J. M. Jornet, Q. H. Abbasi, and A. Alomainy, "A comprehensive survey on hybrid communication in context of molecular communication and terahertz communication for body-centric nanonetworks," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 6, no. 2, pp. 107–133, 2020.

[4] J. T. Gomez *et al.*, "Fine-tuned circuit representation of human vessels through reinforcement learning: A novel digital twin approach for hemodynamics," in *Proceedings of the 10th ACM International Conference on Nanoscale Computing and Communication*, 2023, pp. 46–52.

[5] A. K. Shrivastava *et al.*, "Adaptive threshold detection and isi mitigation in mobile molecular communication," in *2020 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2020, pp. 1–6.

[6] G. H. Alshammri *et al.*, "Adaptive batch training rule-based detection scheme for on-off-keying diffusion-based molecular communications," in *2018 IEEE 13th Nanotechnology Materials and Devices Conference (NMDC)*. IEEE, 2018, pp. 1–4.

[7] M. Movahednasab *et al.*, "Adaptive transmission rate with a fixed threshold decoder for diffusion-based molecular communication," *IEEE Transactions on Communications*, vol. 64, no. 1, pp. 236–248, 2015.

[8] A. Zai *et al.*, *Deep reinforcement learning in action*. Manning Publications, 2020.

[9] J. T. Gómez *et al.*, "Optimizing terahertz communication between nanosensors in the human cardiovascular system and external gateways," *IEEE Communications Letters*, 2023.

[10] C. Funck, F. B. Laun, and A. Wetscherek, "Characterization of the diffusion coefficient of blood," *Magnetic resonance in medicine*, vol. 79, no. 5, pp. 2752–2758, 2018.

[11] T.-Y. Tu and P. C.-P. Chao, "Continuous blood pressure measurement based on a neural network scheme applied with a cuffless sensor," *Microsystem Technologies*, vol. 24, pp. 4539–4549, 2018.

[12] M. P. Mota *et al.*, "Adaptive modulation and coding based on reinforcement learning for 5g networks," in *2019 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2019, pp. 1–6.

[13] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.