Leveraging Symmetry in RL-based Legged Locomotion Control

Zhi Su*,2, Xiaoyu Huang*,1, Daniel Ordoñez-Apraez3, Yunfei Li2, Zhongyu Li1, Qiayuan Liao1,
Giulio Turrisi3, Massimiliano Pontil3, Claudio Semini3, Yi Wu2,4, Koushil Sreenath1

***Abstract*— Model-free reinforcement learning is a promising approach for autonomously solving challenging robotics control problems, but faces exploration difficulty without information about the robot's morphology. The under-exploration of multiple modalities with symmetric states leads to behaviors that are often unnatural and sub-optimal. This issue becomes particularly pronounced in the context of robotic systems with morphological symmetries, such as legged robots for which the resulting asymmetric and aperiodic behaviors compromise performance, robustness, and transferability to real hardware. To mitigate this challenge, we can leverage symmetry to guide and improve the exploration in policy learning via equivariance / invariance constraints. We investigate the efficacy of two approaches to incorporate symmetry: modifying the network architectures to be strictly equivariant / invariant, and leveraging data augmentation to approximate equivariant / invariant actor-critics. We implement the methods on challenging loco-manipulation and bipedal locomotion tasks and compare with an unconstrained baseline. We find that the strictly equivariant policy consistently outperforms other methods in sample efficiency and task performance in simulation. Additionaly, symmetry-incorporated approaches exhibit better gait quality, higher robustness and can be deployed zero-shot to hardware.**

I. Introduction

The field of robotics has witnessed a surge in the adoption of data-driven reinforcement learning (RL) methods to tackle the control problems of legged locomotion [1], [2], navigation [3], and manipulation [4]. This trend is primarily fueled by the ability of these methods to (i) cope with phenomena that impact the system evolution but are challenging to model analytically, (ii) autonomously acquire control strategies without the need of extensive domain knowledge. However, commonly-used model-free RL methods often treat the robot as a black-box system; by neglecting analytical models of dynamics, they often remain agnostic to the properties of robot's morphology. Furthermore, these methods face exploration difficulties to learn multi-modalities, especially symmetric modalities [5] where the under-exploration of some modes leads to asymmetric behaviors that are often unnatural and sub-optimal. For example, failure to fully capture the two symmetric modalities of bipedal locomotion leads to limping behaviors and reduced control performance, compromising robustness and transferability to real hardware.

Leveraging symmetries in Markov decision processs (MDPs) is a promising direction to alleviate the difficulty in symmetric-modality learning and provides a strong bias that we can leverage to improve the policy's exploration. Specifically, since symmetric MDPs possess equivariant optimal control policies [6], [7], posting equivariance requirements on the RL policy helps it approach such an optimal policy. Since symmetry is ubiquitous in both biological and robotic systems, it is indeed an important but under-explored question of the best way to properly incorporate symmetry into RL algorithms. We note that although equivariance has gained attention in other fields such as computer vision and graphics, there have been limited prior works that apply it on legged robots and demonstrate its efficacy in related tasks.

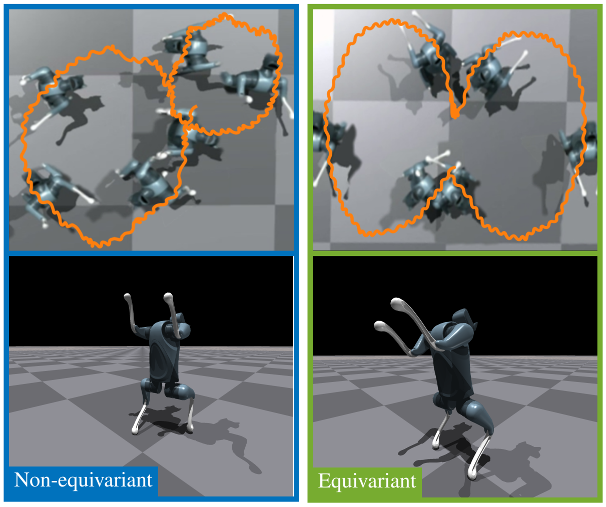

Fig. 1: Comparison between non-equivariant (left) and equivariant (right) control policies, of a quadrupedal robot with a sagittal reflection symmetry, performing a right / left bipedal turning task. Top plots show the trajectories of a commanded turn in opposite directions. Bottom figures visualize the gait pattern the policy learns. While the unconstrained policy learns an asymmetric gait between left and right feet and fails to perform a symmetric turning trajectory, the equivariant policy achieves both motion-level and task-level symmetry. For more experimental results, readers are encouraged to check out the supplementary video at `https://youtu.be/Ad1clt4Yi4U` or project website at `https://suz-tsinghua.github.io/symmloco/`.

In this work, we investigate different methods for adding equivariance information in improving the exploration guidance in training model-free RL algorithms. Specifically, we investigate how a loosely equivariant policy trained by data augmentation as carried out in [8] compare to a strictly equivariant policy enforced by its network architecture on legged robot control. We perform extensive experiments in the challenging tasks of loco-manipulation and bipedal locomotion as quadrupeds. We benchmark the two methods against a vanilla RL policy in simulation environments and showcase the optimality of the symmetry-incorporated

*Equal contribution. [1]UC Berkeley, CA, USA. [2]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China. [3]Istituto Italiano di Tecnologia, Italy. [4]Shanghai Qi Zhi Institute, Shanghai, China. Contact: haytham.huang@berkeley.edu

policies. Our experimental results show that the equivariance-enforced policy consistently outperforms other variants both in terms of the performance metrics and the acquired gait patterns being more steady and natural.

We demonstrate the sim-to-real capability of both methods for completing a loco-manipulation task and a bipedal locomotion task, and show the enhanced robustness of the symmetry-incorporated policies compared to a vanilla RL policy. We provide a detailed discussion on the robustness of the two variants on real-world hardware, which could provide guidance for future development using symmetry-incorporated RL methods for different tasks.

II. Incorporating Symmetry in Model-free RL

We introduce two variations of PPO [9] leveraging the robot's morphological symmetry: PPO with data augmentation in training (**PPOaug**) and PPO with hard equivariance/invariance constraints on neural networks (**PPOsymm**). These are compared with a vanilla PPO implementation. Symmetric loss function methods were excluded as they were outperformed by PPOaug [8]. For symmetry representations ρ_S and ρ_A, we use MorphoSymm [10] and ESCNN [11].

A. PPOaug

This method leverages data augmentation on policy π_θ and critic V_ϕ by using both augmented and collected transition tuples. Data augmentation biases the critic to be approximately $\mathbb{G}$-invariant, guiding the actor towards an invariant return estimate. Since PPO is on-policy, augmented samples introduce off-policy effects, which we mitigate by ensuring π_θ is approximately equivariant. To achieve this, we minimize gradient differences between original and augmented samples by performing augmentation over each mini-batch during policy updates, and initialize the policy network with zero mean and small variance. This ensures the policy remains approximately equivariant after updates.

B. PPOsymm

To enforce $\mathbb{G}$-equivariance on π_θ and $\mathbb{G}$-invariance on V_ϕ, we use equivariant/invariant neural networks, specifically EMLP, which ensures $\pi_\theta(g \triangleright \boldsymbol{s}) = g \triangleright \pi_\theta(\boldsymbol{s})$. The invariant critic is designed by substituting the final layer of the EMLP with an invariant transformation. This makes $V_\phi(g \triangleright \boldsymbol{s}) = V_\phi(\boldsymbol{s})$. Additionally, we consider temporal symmetry by assuming gait periodicity, allowing valid equivariance transformation on the phase signal ψ. Including ψ helps avoid the problem of neutral states where $g \triangleright \boldsymbol{s} = \boldsymbol{s}$.

III. Tasks

We compare the Proximal Policy Optimization (PPO) variants on four tasks involving loco-manipulation and bipedal locomotion in quadrupedal robots. Observations include proprioceptive and task-specific data. All rewards are $\mathbb{G}$-invariant. Training is conducted in the Isaac Gym [12] with domain randomization.

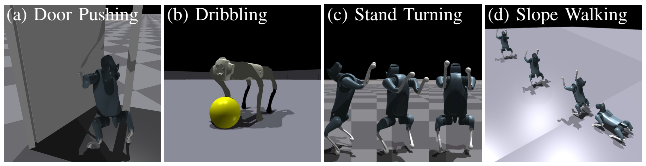

Fig. 2: Challenging tasks to test the efficacy of symmetry-incorporated policies. (a,b) showcase loco-manipulation with task-space symmetry and (c,d) exhibit bipedal locomotion with motion-level symmetry. These tasks push the boundary of agility and dynamic mobile manipulation for quadrupeds.

Door Pushing: In this task (Fig. 2(a)), the robot stands and opens a door using its front limbs, adjusting for doors that swing left or right. Observations include the door's relative position, orientation, and swing direction. The reward involves forward velocity tracking to encourage door-pushing.

Dribbling: Shown in Fig. 2(b), the robot dribbles a soccer ball based on desired velocity commands. Observations include the ball's relative position to the robot, and rewards focus on ball velocity tracking and proximity to the ball [13].

Stand Turning: This agile bipedal locomotion task, adapted from [14], requires the robot to stand on two feet and follow input commands for linear velocity and heading, with rewards based on tracking accuracy.

Slope Walking: In this task, the robot traverses an inclined surface on two feet, using a terrain curriculum with a maximum slope of $11.3°$.

IV. Training and Evaluation

In this section, we discuss the training and evaluation of PPO, PPOaug, and PPOeqic across the described tasks. The training scheme remains consistent for all methods, with separate hyperparameter tuning for each task. We evaluate performance on 2,000 simulation environments, with results averaged over three different seeds.

A. Training Performance

We assess training return and sample efficiency as the criteria for training performance. As shown in Fig. 3, PPOeqic consistently outperforms other variants, demonstrating the effectiveness of equivariance constraints. PPOeqic also exhibits improved sample efficiency, especially in early training stages, indicating more efficient exploration guidance. PPOaug achieves high training returns but similar sample efficiency compared to vanilla PPO. Unlike PPOeqic, PPOaug must learn two approximately equivariant action distributions, making it a more complex learning process. Vanilla PPO often converges to suboptimal policies that overfit to simulation, lacking robustness for sim-to-real transfer.

B. Door Pushing Task

The door-pushing task benchmarks task-level symmetry in loco-manipulation using success rate (SR) and the Symmetric Index (SI) [15]. Note that a higher SI indicates a more *asymmetric* behavior, which is not desired.

1) Success Rate: As shown in Table I, PPOeqic achieves a 4.47% higher mean SR than PPO, while PPOaug has a 6% lower mean SR. Both symmetry-incorporated policies reach a higher maximum SR, around 88% for PPOeqic and 87% for PPOaug, suggesting a greater likelihood of optimal performance. PPOeqic also shows a 10% higher mean SR

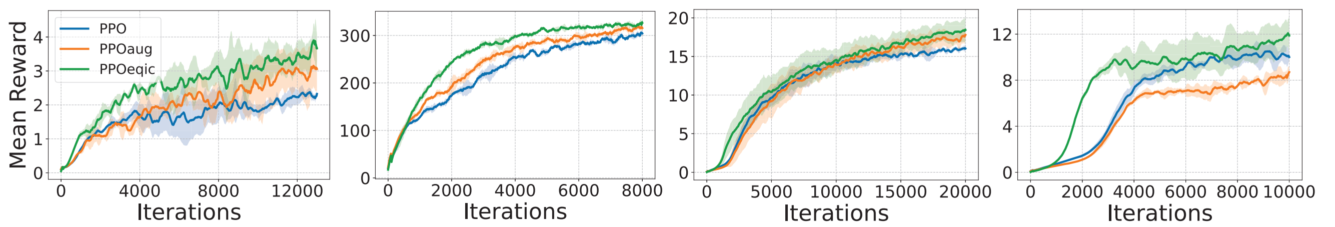

Fig. 3: Comparison of training curves of PPO, PPOaug, and PPOeqic on four tasks from left to right: Door Pushing, Dribbling, Stand Turning, and Slope Walking. Learning curves show mean episodic return and standard deviation for three seeds. PPOeqic consistently demonstrates the highest training returns and sample efficiency in all tasks.

Method		**Mean SR** (%)	**Max SR** (%)	**RSI**	**OOD Mean SR** (%)	**OOD Max SR** (%)	**OOD RSI**
PPO	trained on 1 side	43.40 ± 1.73	44.47	199.96	27.46 ± 2.34	30.04	199.99
	trained on 2 sides	61.18 ± 7.56	69.63	12.25	42.98 ± 2.21	45.51	3.45
PPOaug	trained on 1 side	54.39 ± 32.56	86.98	3.02	38.24 ± 19.17	52.85	1.40
	trained on 2 sides	50.24 ± 36.52	74.98	3.77	36.46 ± 25.98	53.78	**0.17**
PPOeqic	trained on 1 side	**65.65** ± 23.16	**87.96**	**0.98**	**44.15** ± 9.39	50.63	0.88
	trained on 2 sides	59.92 ± 17.39	74.74	1.65	38.56 ± 15.54	**55.95**	0.32

TABLE I: Comparison of success rates (SR) and their symmetry index on door-pushing tasks on training-distribution and out-of-distribution scenarios. Of the three variants, PPOeqic demonstrates both higher success rate and better symmetry index in both cases, indicating a better task-level symmetric policy.

than PPOaug, indicating that training can be more robust over different seeds with hard-constraint equivariance compared to that induced by symmetric augmentation.

2) Symmetry Index: Shown in Table I, PPO receives a high SI, 4 times higher than PPOaug and 8 times higher than PPOeqic, indicating its failure to learn balanced policies between symmetric modes. In contrast, PPOaug and PPOeqic achieve low SI values, demonstrating symmetric behavior in the task space.

3) Out-of-Distribution Scenarios: In out-of-distribution scenarios, PPOeqic outperforms PPO in both mean and maximum SRs. It demonstrates more stable performance than PPOaug, achieving a 5.9% higher mean SR across three seeds, showing better generalization to symmetric MDPs.

4) Single Mode Training: Interestingly, PPOeqic trained only on one side scenario consistently outperforms training on both sides. This trend also appears in PPOaug, despite the lack of enforced equivariance. We attribute this to domain randomization of joint properties, which breaks the assumption of $\mathbb{G}$-invariant state transition probability. In cases of asymmetrical randomization, training on only one mode allows the policy to leverage symmetry effectively. It can mirror an equivariant transition without conflicting with the training data, forming a symmetric MDP. However, if both modes are included, mirrored transitions in the training data can cause conflicts due to asymmetric dynamics. Furthermore, PPOeqic trained with right-handed doors can be zero-shot deployed on doors swinging either way.

C. Dribbling Task

Generally, symmetry improves dribbling gaits. PPOeqic and PPOaug keep the ball closer to the robot, while PPO often kicks it away. Quantitatively, PPOeqic achieves a slight edge in episodic return, suggesting near-saturation in the simulation environment for all methods.

Method	**Error (rad)**	**Error RSI**	**CoT (Js/m)**	**CoT RSI**
PPO	0.265 ± 0.022	0.0945	2223 ± 102	0.0259
PPOaug	0.259 ± 0.010	0.0212	2378 ± 101	**0.0046**
PPOeqic	**0.254** ± 0.014	**0.0207**	**2026** ± 187	0.0172

TABLE II: Comparison of command tracking error, Cost of Transport and their symmetry index on stand turning tasks for three PPO variants. PPOeqic demonstrates less error and energy consumption, indicating a more optimal policy.

D. Stand Turning Task

In this bipedal locomotion task, PPO develops a staggered gait that results in jittering and lacks symmetric turning. In contrast, symmetry-incorporated policies exhibit both symmetric gaits and turning trajectories. PPOeqic achieves lower tracking error and cost of transport (CoT) compared to other variants, indicating relative optimality. PPOaug shows lower tracking error and error SI but higher CoT than PPO, suggesting less optimal behavior than PPOeqic.

E. Slope Walking Task

In the slope walking task, we observe significant differences between the three PPO variants. While the PPO policy learns a relatively natural gait, it struggles with symmetric foot placement, leading to backward steps and frequent balance loss, as shown in Fig. 4(a). This results in poor velocity tracking, with the policy covering only half the distance of PPOeqic in the same timeframe. The PPOaug policy shows improvement, with better alternation between feet, but still exhibits variations in step size and occasional staggering (Fig. 4(b)). PPOeqic demonstrates the most stable gait, with consistent foot exchange and regulated contact sequences on an $11.3°$ incline (Fig. 4(c)). It achieves the desired walking velocity of 0.25 m/s, outperforming PPOaug by 20%.

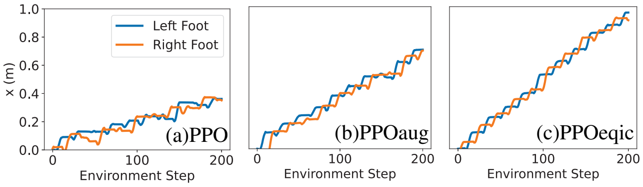

Fig. 4: Plots of the feet positions in the desired walk direction. We observe that vanilla PPO learns an unstable step pattern with backward steps and foot slipping, resulting in 50% slower walking speed. PPOaug improves drastically but asymmetric patterns such as foot dragging still exists. PPOeqic presents the most symmetric interweaving gait pattern and walks at the desired speed.

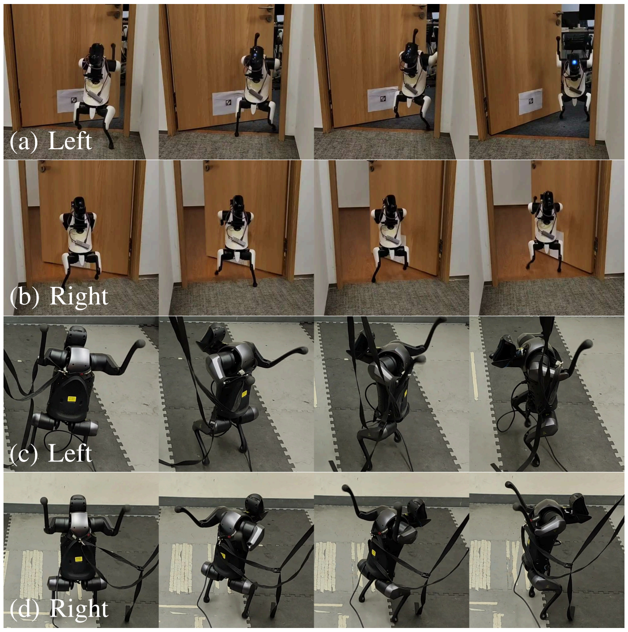

Fig. 5: Snapshots of equivariant policies deployed zero-shot to perform real-world tasks of door pushing (a,b) and stand turning (c,d). For task-level symmetry (e.g. door pushing), PPOeqic trained with only right-handed door can be deployed on both left and right-handed doors with symmetric gait patterns. For motion symmetry, PPOaug is more robust against slightly asymmetric dynamics that exists on actual hardware.

F. *Real-world Experiments*

We assess sim-to-real transfer through real-world experiments on two selected tasks.

1) Door Pushing Task: Although vanilla PPO achieves high success rates in simulation, it underperforms in real-world scenarios, with the asymmetric gaits often causing the robot to fall due to overfitting to simulated contact dynamics. In contrast, PPOeqic demonstrates greater robustness; even when trained on right-handed doors, it can effectively open both left and right-handed doors. PPOaug, however, tends to rely on one front leg, showing limited pushing strength. This underscores the value of equivariance constraints for achieving optimal real-world performance.

2) Stand Turning Task: PPO fails in real-world settings, while PPOaug and PPOeqic demonstrate good zero-shot transfer, executing consistent 90-degree turns. PPOaug exhibits higher robustness, completing 9 out of 10 trials. This again highlights the substantial enhancement in sim-to-real transfer by incorporating symmetry into the learning process.

3) Discussion: As expected, real-world environments are not perfectly symmetric MDPs, attributed to the asymmetries resulted from hardware, ground conditions, and other complicated factors. PPOeqic excels in tasks dominated by task-space symmetry where intrinsic asymmetries are negligible. In tasks with intrinsic symmetry, PPOeqic is more sensitive to distribution shifts. PPOaug allows adaptation to the robot's asymmetries, enhancing robustness against imperfect symmetry. The choice between PPOaug and PPOeqic should be tailored to the problem's nuances.

V. Conclusion

In this work, we investigate the benefits of leveraging symmetry in model-free RL for legged locomotion. We compare the performance of incorporating data augmentation and hard equivariance constraints across four challenging bipedal locomotion and loco-manipulation tasks against a vanilla PPO baseline. We find that imposing hard symmetry constraints on the learned policy and value functions leads to better performance than other methods. Compared to vanilla PPO, equivariant policies learn notably more steady and symmetric gait patterns, eventually leading to better task-space symmetry. More importantly, equivariant policies trained on single symmetry mode are directly generalizable to other modes. When applied to real-world scenarios, symmetry-incorporated policies demonstrate significantly better robustness than unconstrained policy. Furthermore, PPOaug copes with slight asymmetry in robot's own dynamics, while PPOeqic demonstrates better performance on task-space symmetry.

We hope this work can aspire the exploration of leveraging symmetry constraints on robots with larger symmetry groups than the reflection group concerned in this work. As the number of symmetry modalities increases, the symmetry constraints are expected to play a more crucial role in guiding the exploration of model-free RL over the increasingly complex state, action spaces. In addition, equivariant policies demonstrate promising potential for even larger performance gains over vanilla PPO, highlighting improvements as large as 26% in some seeds. Future efforts could be on stabilizing training to consolidate this enhancement and develop better symmetry-incorporated RL algorithms.

References

[1] X. B. Peng, E. Coumans, T. Zhang, T.-W. E. Lee, J. Tan, and S. Levine, "Learning agile robotic locomotion skills by imitating animals," in *Robotics: Science and Systems*, 2020.

[2] Z. Li, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, "Reinforcement learning for versatile, dynamic, and robust bipedal locomotion control," *arXiv preprint arXiv:2401.16889*, 2024.

[3] E. Wijmans, A. Kadian, A. Morcos, S. Lee, I. Essa, D. Parikh, M. Savva, and D. Batra, "Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames," in *International Conference on Learning Representations*, 2019.

[4] A. Singh, L. Yang, C. Finn, and S. Levine, "End-to-end robotic reinforcement learning without reward engineering," *Robotics: Science and Systems XV*, 2019.

[5] L. Lee, B. Eysenbach, E. Parisotto, E. Xing, S. Levine, and R. Salakhutdinov, "Efficient exploration via state marginal matching," *arXiv preprint arXiv:1906.05274*, 2019.

[6] D. Wang, M. Jia, X. Zhu, R. Walters, and R. Platt, "On-robot learning with equivariant models," in *6th Annual Conference on Robot Learning*, 2022.

[7] M. Zinkevich and T. Balch, "Symmetry in markov decision processes and its implications for single agent and multi agent learning." Citeseer, 2001.

[8] M. Mittal, N. Rudin, V. Klemm, A. Allshire, and M. Hutter, "Symmetry considerations for learning task symmetric robot policies," *arXiv preprint arXiv:2403.04359*, 2024.

[9] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[10] D. Ordoñez-Apraez, G. Turrisi, V. Kostic, M. Martin, A. Agudo, F. Moreno-Noguer, M. Pontil, C. Semini, and C. Mastalli, "Morphological symmetries in robotics," *The International Journal of Robotics Research*, 2024.

[11] G. Cesa, L. Lang, and M. Weiler, "A program to build e (n)-equivariant steerable cnns," in *International conference on learning representations*, 2021.

[12] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa *et al.*, "Isaac gym: High performance gpu based physics simulation for robot learning," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[13] Y. Ji, G. B. Margolis, and P. Agrawal, "Dribblebot: Dynamic legged manipulation in the wild," *International Conference on Robotics and Automation*, 2023.

[14] Y. Li, J. Li, W. Fu, and Y. Wu, "Learning agile bipedal motions on a quadrupedal robot," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[15] R. Robinson, W. Herzog, and B. M. Nigg, "Use of force platform variables to quantify the effects of chiropractic manipulation on gait symmetry." *Journal of manipulative and physiological therapeutics*, 1987.