

HeFDI Data Week 2025



Date	Topic	Presenter
27. June 2025	From Simulation to Prediction: AI and Synthetic Data in Biomedicine	Silvia Becker (Institute of Biomedical Engineering (IBT))

Abstract:

Synthetic data, such as electrocardiograms (ECGs) generated through electrophysiological simulations, can be used as a powerful tool to overcome the challenges of accessing large, high-quality datasets in healthcare, which are often difficult to obtain due to privacy concerns, variability and challenges in annotation. This talk will look at how synthetic data can complement real-world datasets, improve model generalisation, facilitate learning in rare or underrepresented scenarios as well as improve the accuracy and reliability of machine learning models in biomedical applications. Further, examples of using such data to train deep learning models for various biomedical applications will be discussed together with potentials and key challenges. Ultimately, the talk explores how synthetic data can help to bridge the gap between computational modeling and real-world clinical AI applications.

About the HeFDI Data Week:

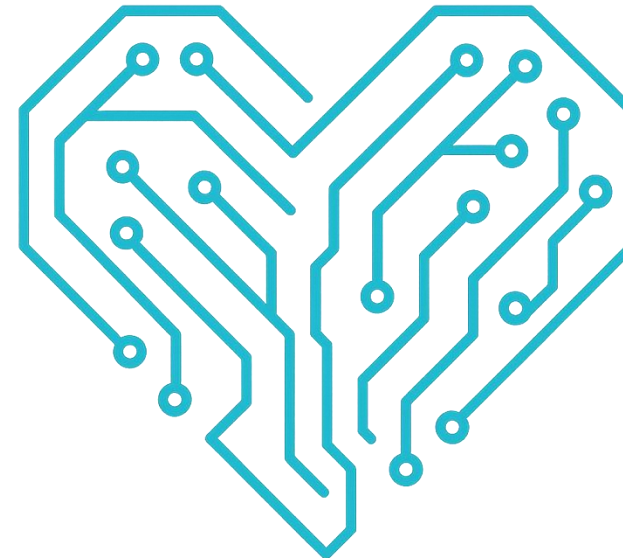
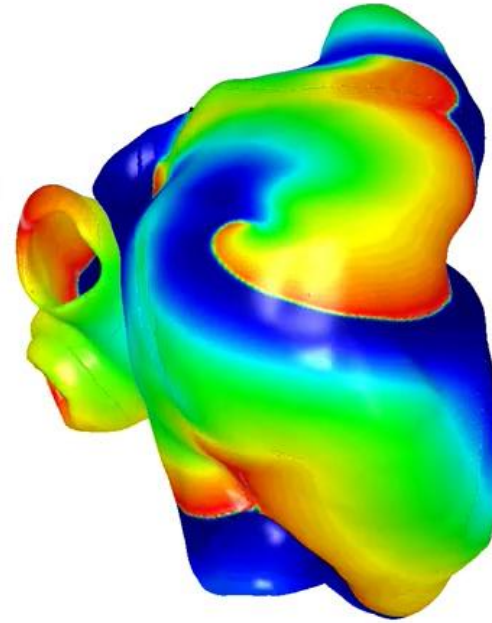
The HeFDI Data Week 2025 is a multi-day online event series which is offered in the context of the nationwide "Digitaltag". The series is aimed at researchers, teachers, students and anyone who wants to learn more about data management, FAIR data and code. Over the course of the week, various topics, developments and challenges related to research data will be covered, such as tools and offers for disciplines from NFDI consortia, legal aspects of research data management as well as research data management and artificial intelligence. The HeFDI Data Week is an offer of the federal state initiative HeFDI - Hessian Research Data Infrastructures, which is funded by the Hessian Ministry of Higher Education, Research, Science and the Arts (HMWK).

DOI-Link: <https://doi.org/10.5281/zenodo.15756218>. Licence information: Creative Commons Attribution 4.0 International ([CC BY 4.0](#))



From Simulation to Prediction: AI and Synthetic Data in Biomedicine

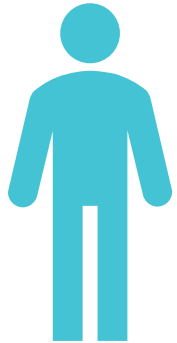
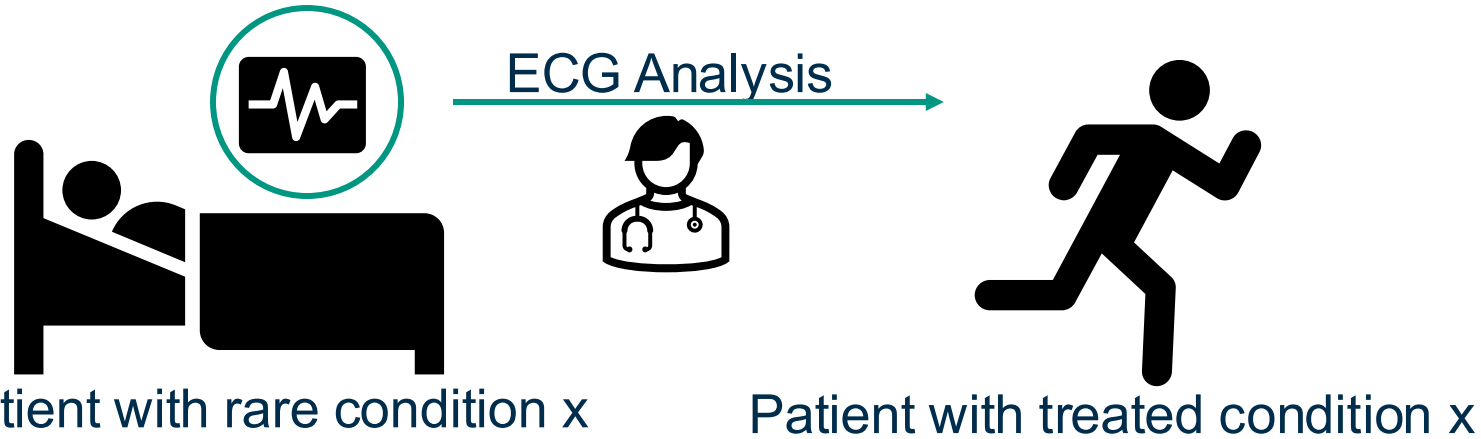
Silvia Becker
Institute of Biomedical Engineering (IBT)



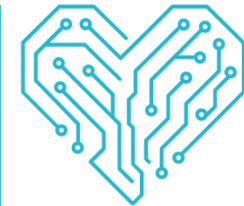
Content

1. Opening Example 
2. Challenges of AI in Biomedicine 
3. Background 
 - Cardiac Modeling
 - Digital Twins & Chimera
4. Solution: Synthetic Data 
 - Potentials
 - Challenges
5. Clinical Applications 
6. Take Home Messages 

Opening Example



**Can we also diagnose this person?
How can AI help us?**



Challenges of AI in Biomedicine

- Missing ground truth datasets with adequate size and high quality labels
 - Big data but small data
- Biased, unbalanced datasets
 - Do we capture the relevant range?
 - Fairness
- Intransparent methodology
 - FAIRness, partly due to data privacy
 - Explainability



F indable



A ccessible



I nteroperable



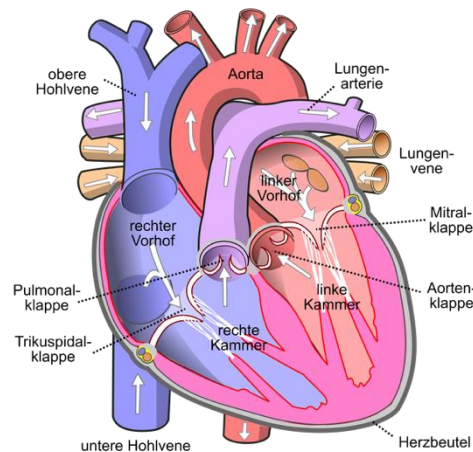
R eusable

www.biomeris.it



Challenges of AI in Biomedicine

- Atrial fibrillation (AF) is the most common sustained arrhythmia



Jakov, CC BY-SA 3.0

Healthy case:
Sinus rhythm



J. Heuser, CC BY-SA 3.0

Pathology:
Atrial fibrillation



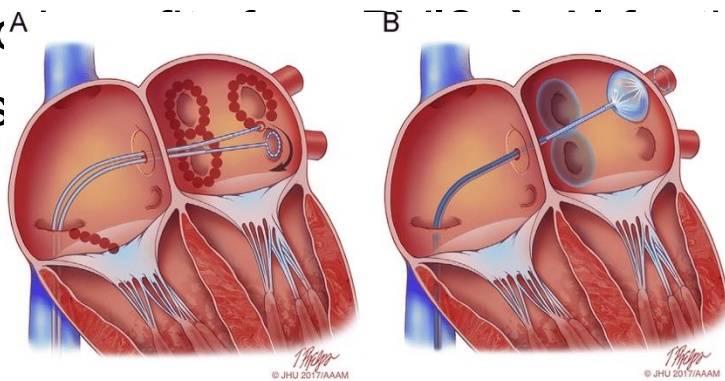
J. Heuser, CC BY-SA 3.0

Challenges of AI in Biomedicine

- Ablation is a standard procedure to treat AF
 - Specifically pulmonary vein isolation (PVI)
 - But: high recurrence rates after PVI

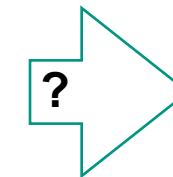
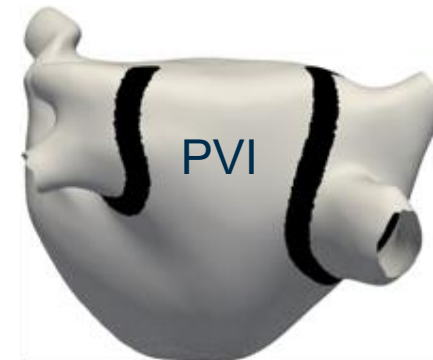
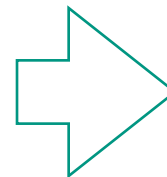
■ What is the rescue?

■ Is

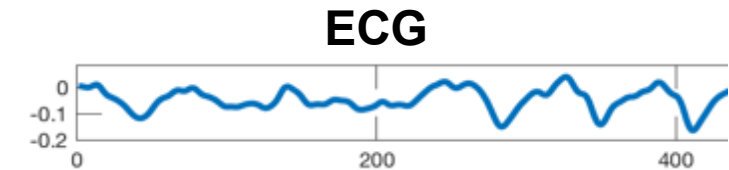


Hopkins, CC BY-NC-ND

Atrial
fibrillation

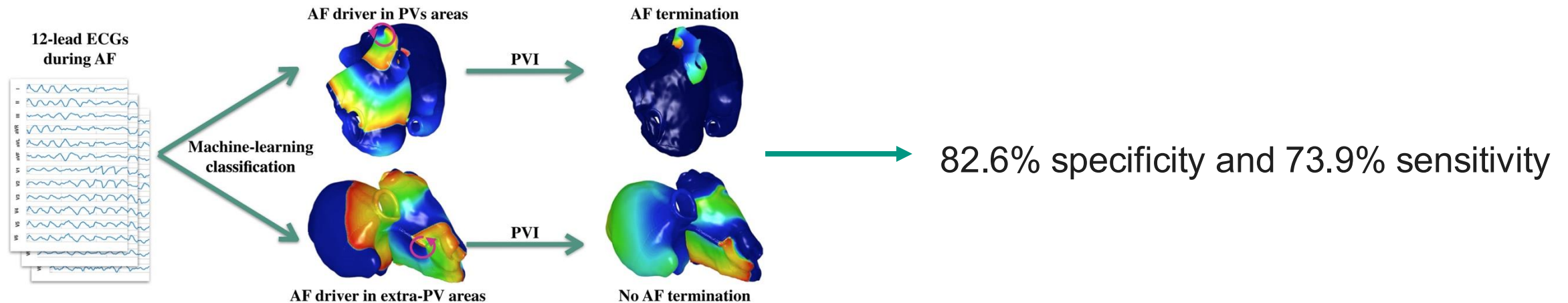
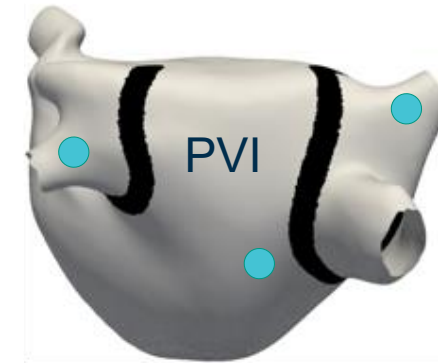


Atrial
fibrillation

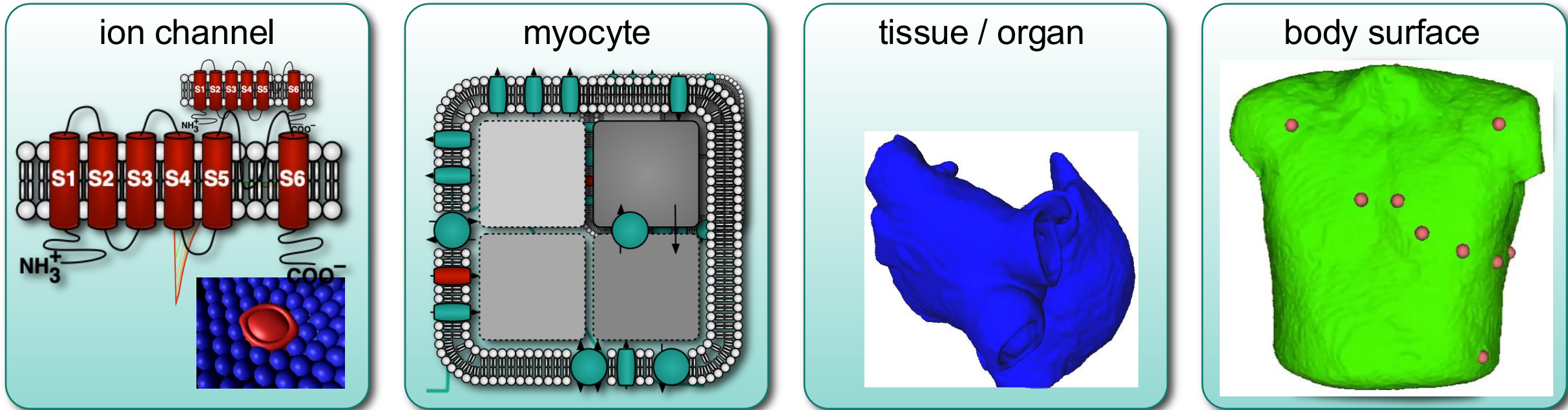


Challenges of AI in Biomedicine

- Simulations to create 1,128 synthetic ECGs
 - Simulations with **AF drivers** in different locations
 - Ground truth available (**within PVs** = PVI success)
- Classifier was trained on this synthetic data
- And tested on a clinical dataset



Cardiac Modeling



biological level of integration

nanometer

micrometer

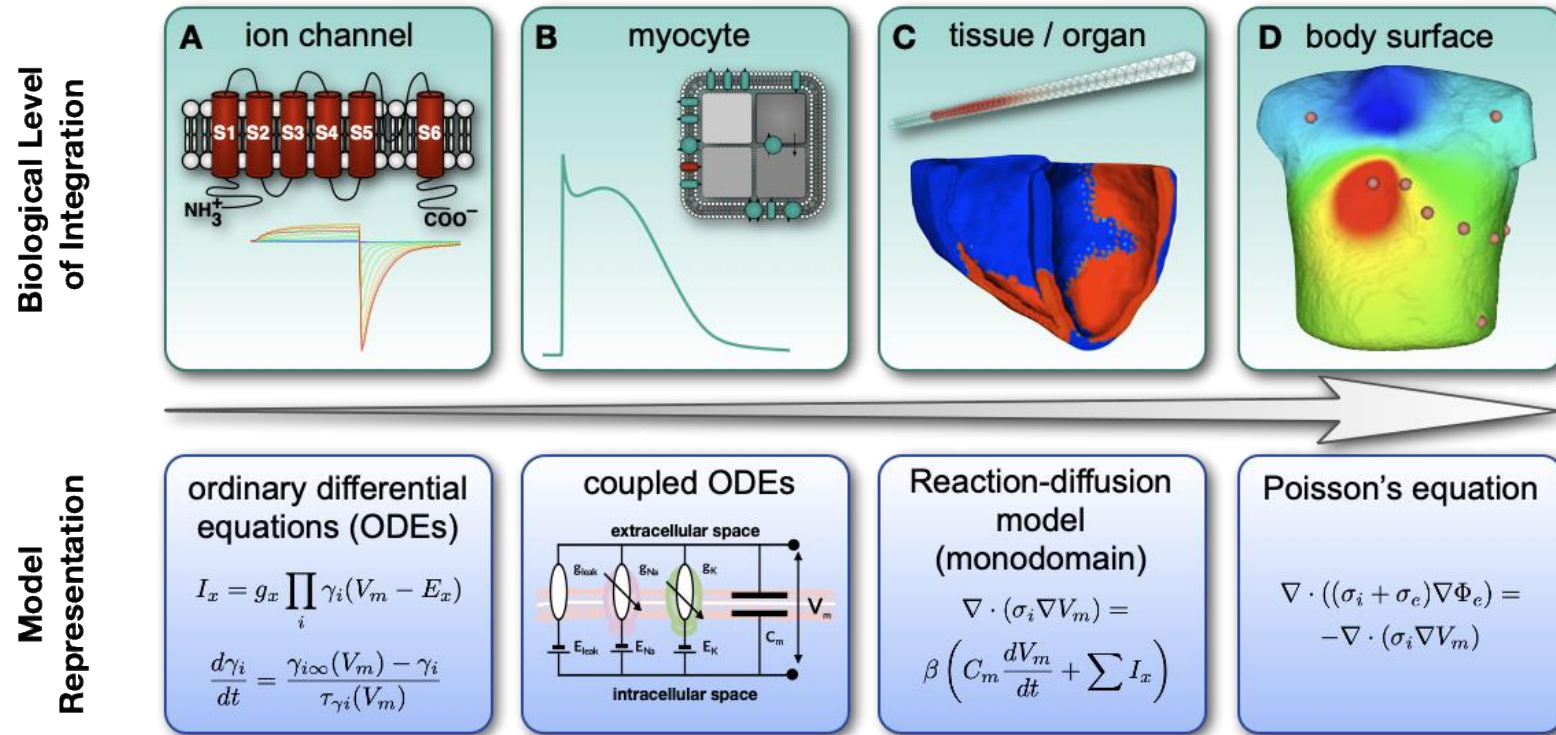
centimeter



Cardiac Modeling



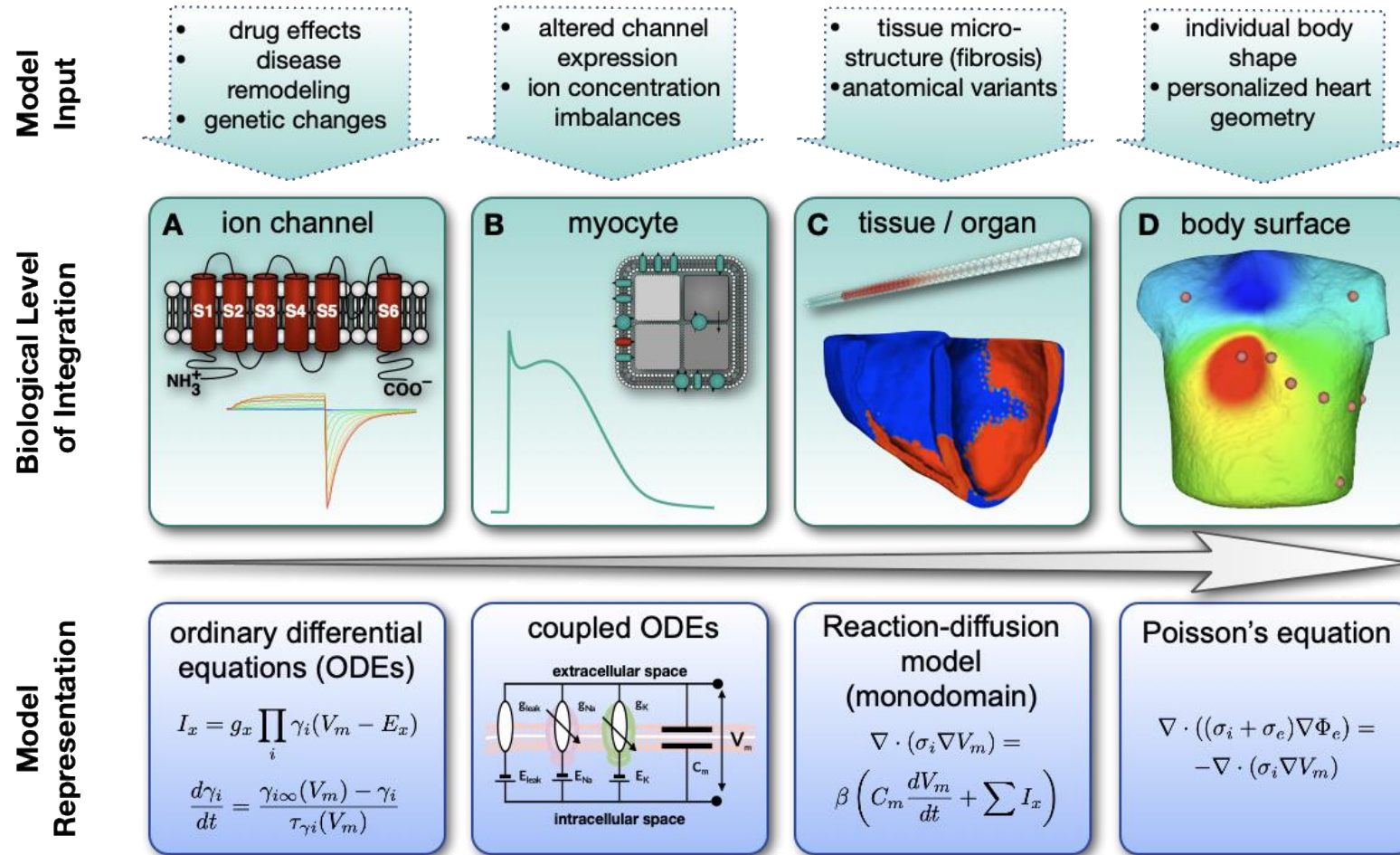
Plank et al., 2021,
Comput Methods Programs Biomed



Cardiac Modeling



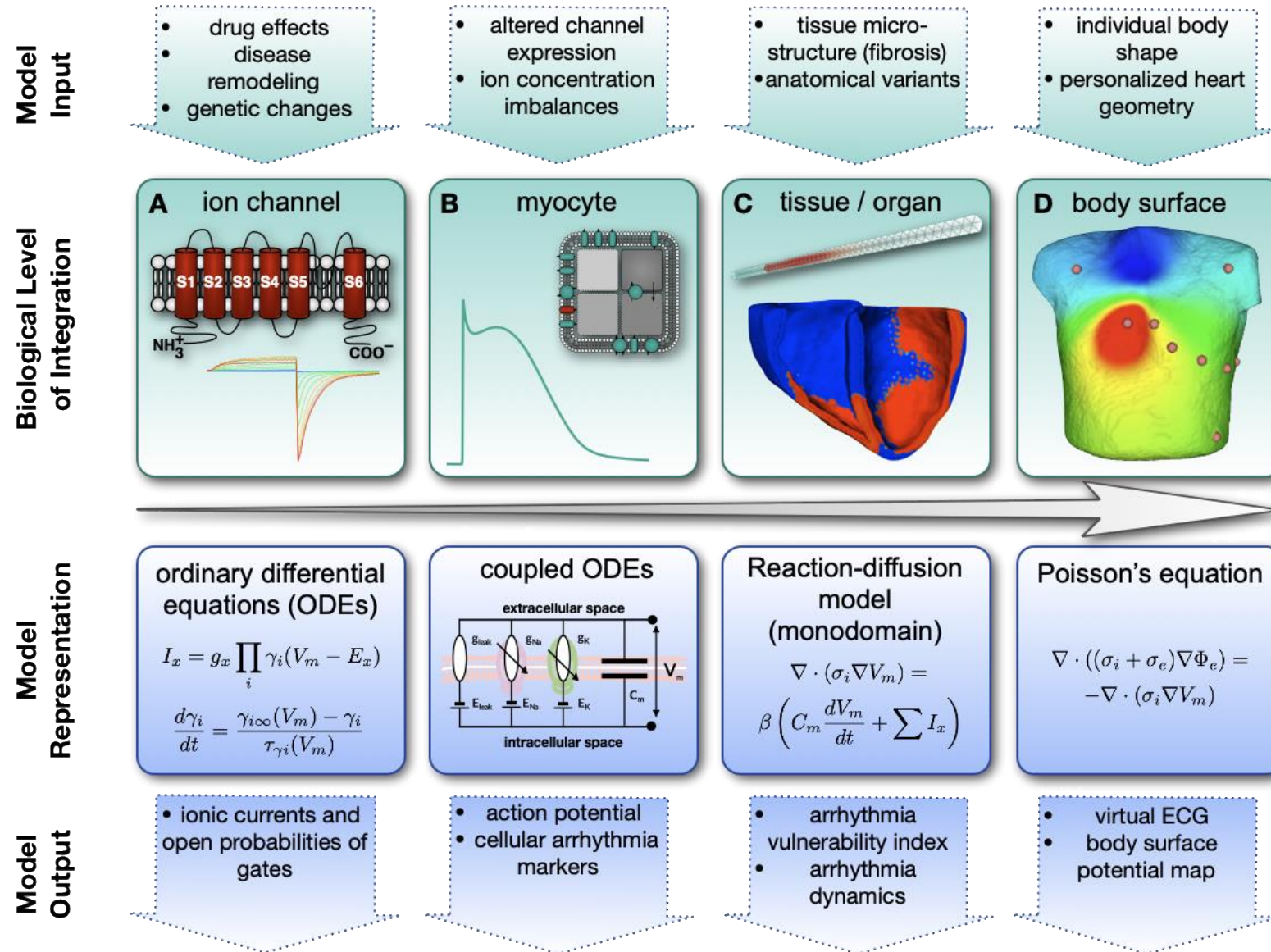
Plank et al., 2021,
Comput Methods Programs Biomed



Cardiac Modeling



Plank et al., 2021,
Comput Methods Programs Biomed

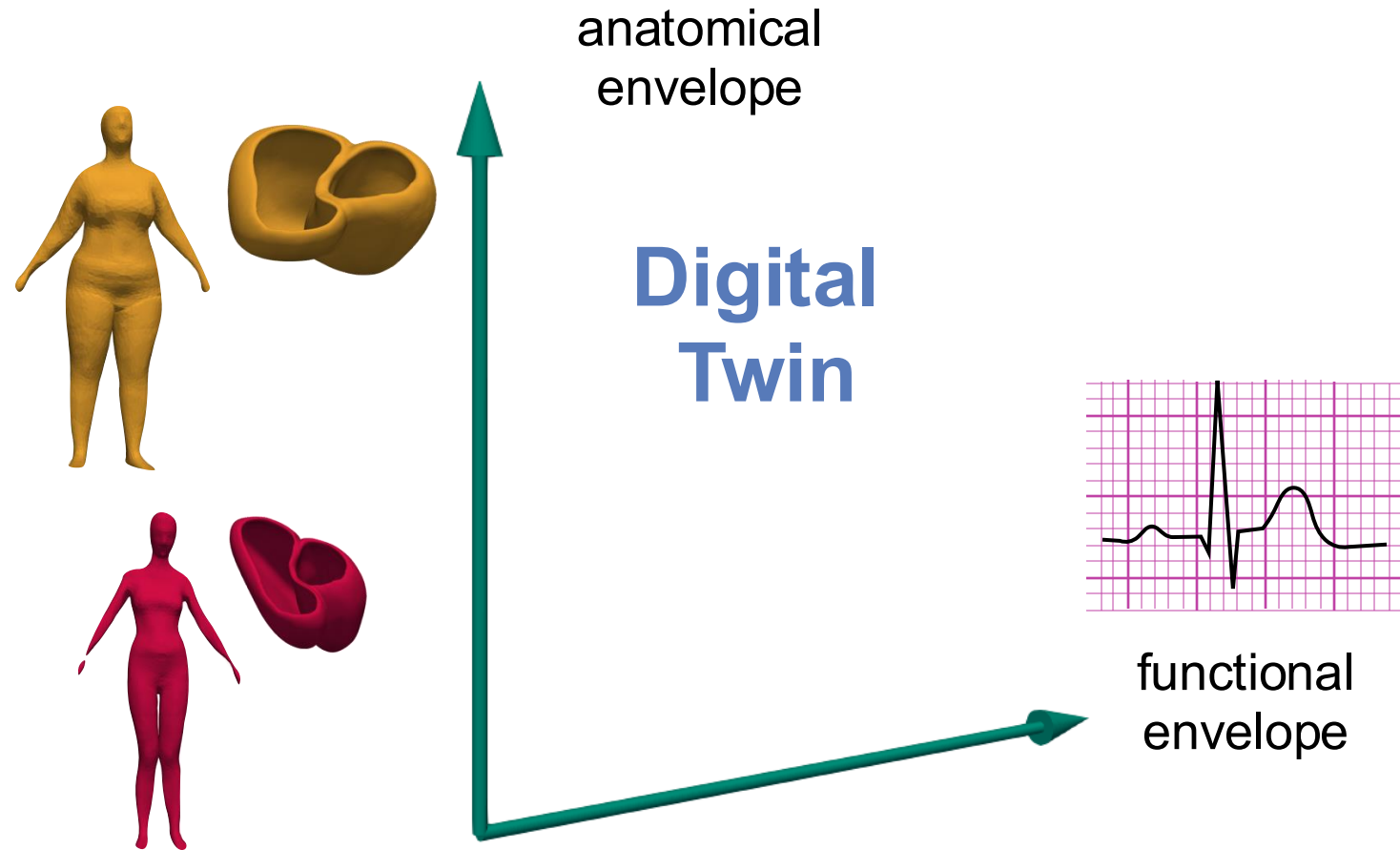


Cardiac Modeling

- Controlled research environment
 - to understand fundamental physiology and pathomechanisms
 - to evaluate diagnostic and therapeutic approaches (in silico studies)
- Scalable research environment to generate big, quality-controlled datasets for machine learning
- Personalized medicine through digital twins

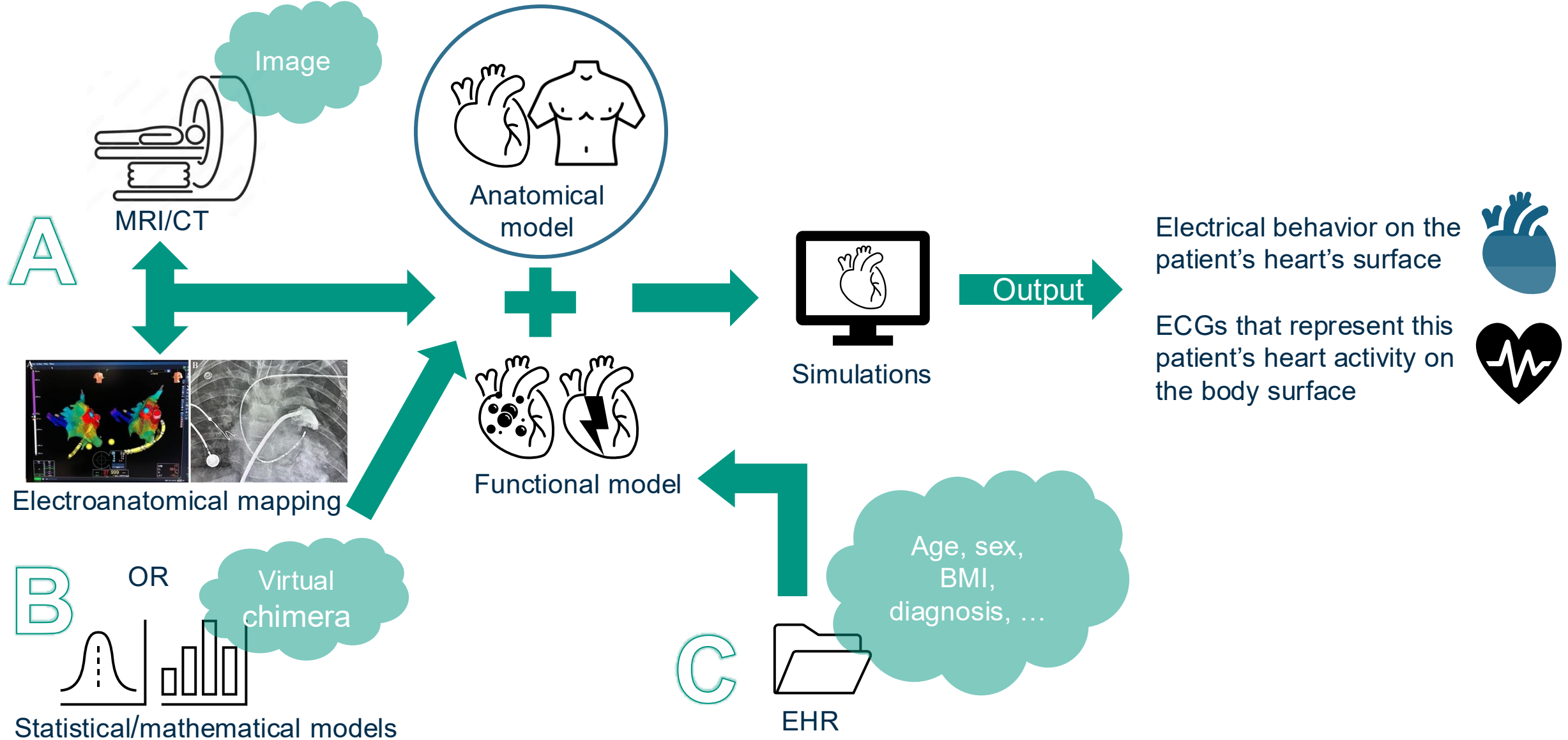


Digital Twins

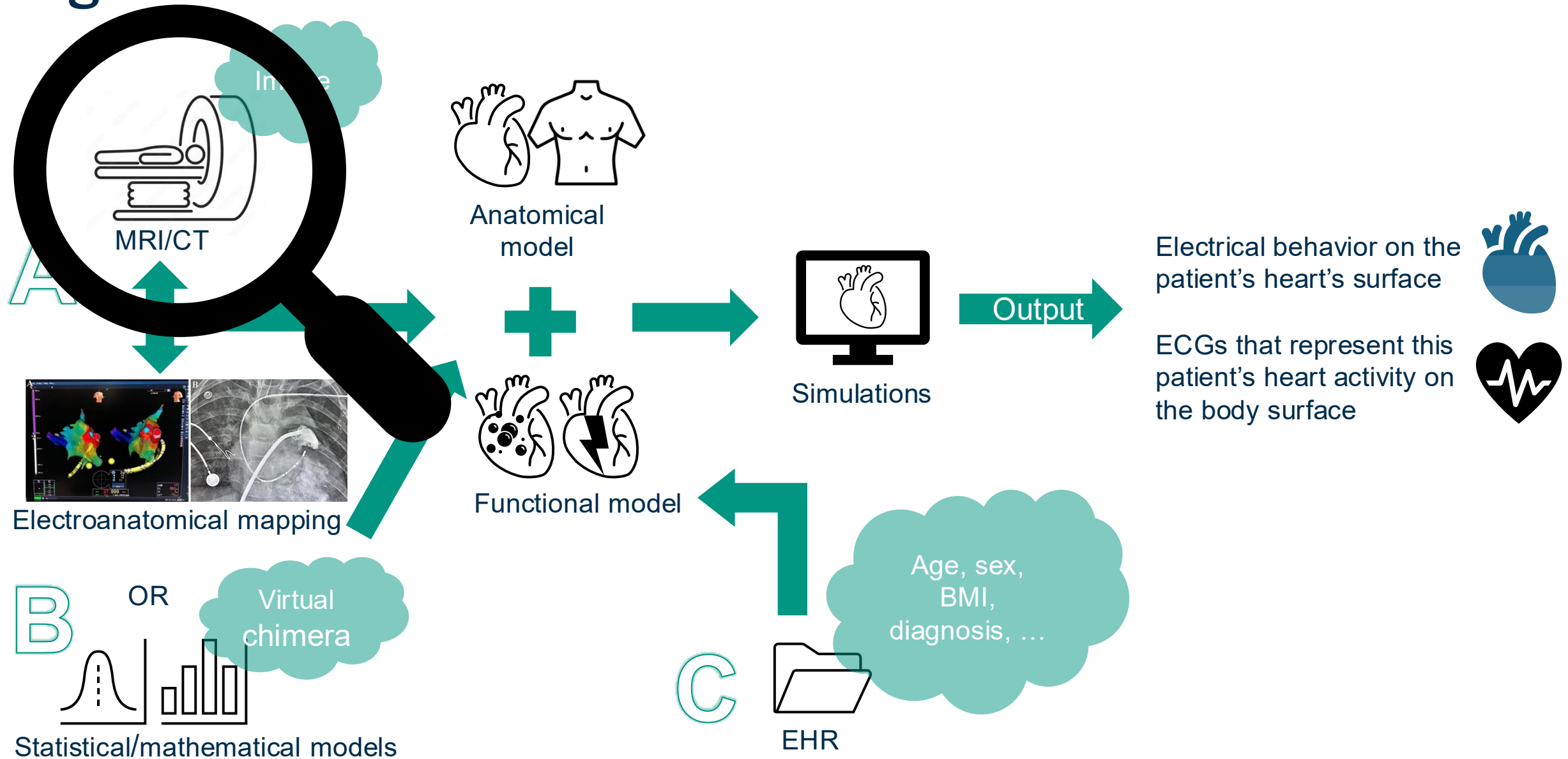


Adapted from Zeike Taylor, University of Leeds

Digital Twins



Digital Twins

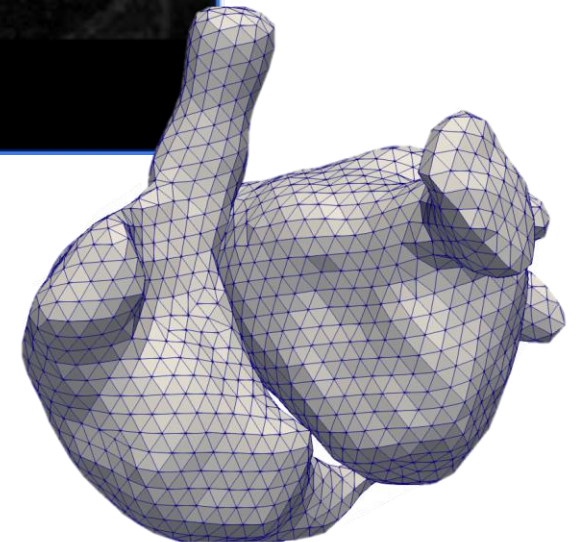
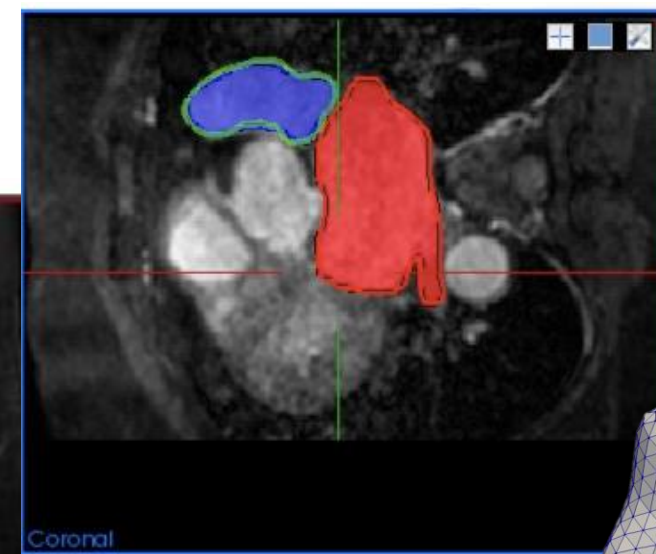
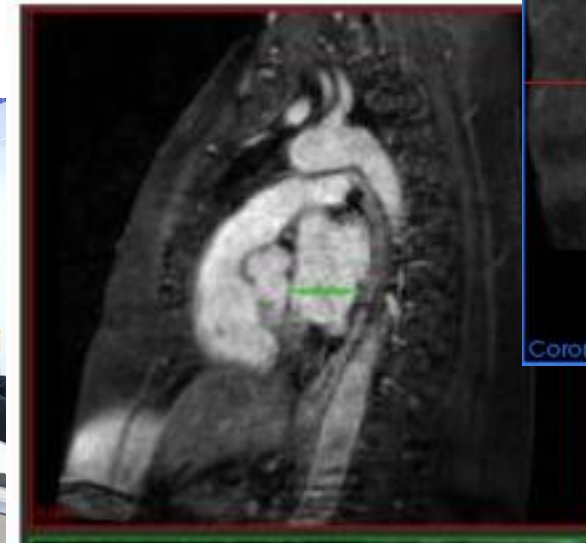


Digital Twins

■ Anatomical personalization



radiologie.de

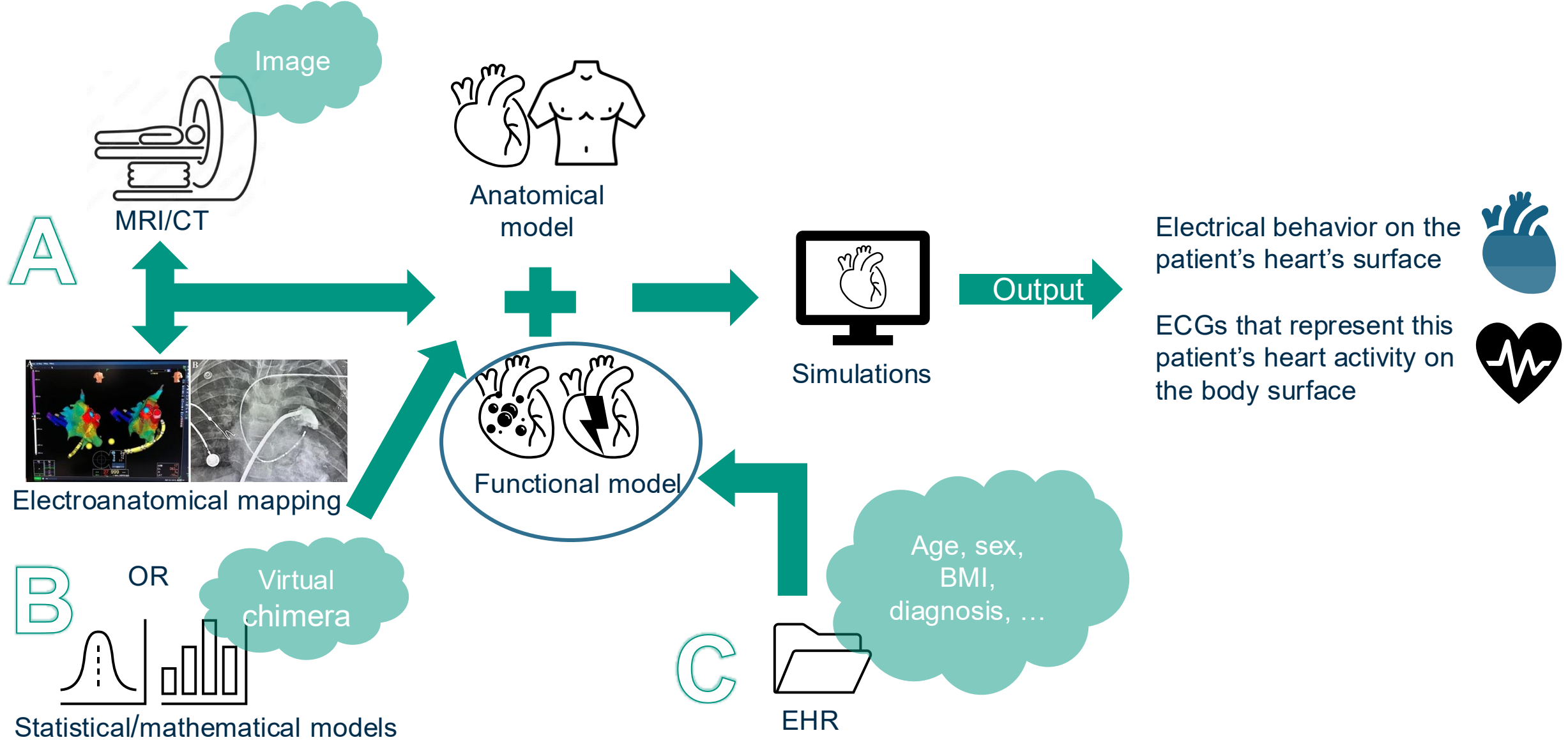


Nagel et al., 2021, Medical Image Analysis

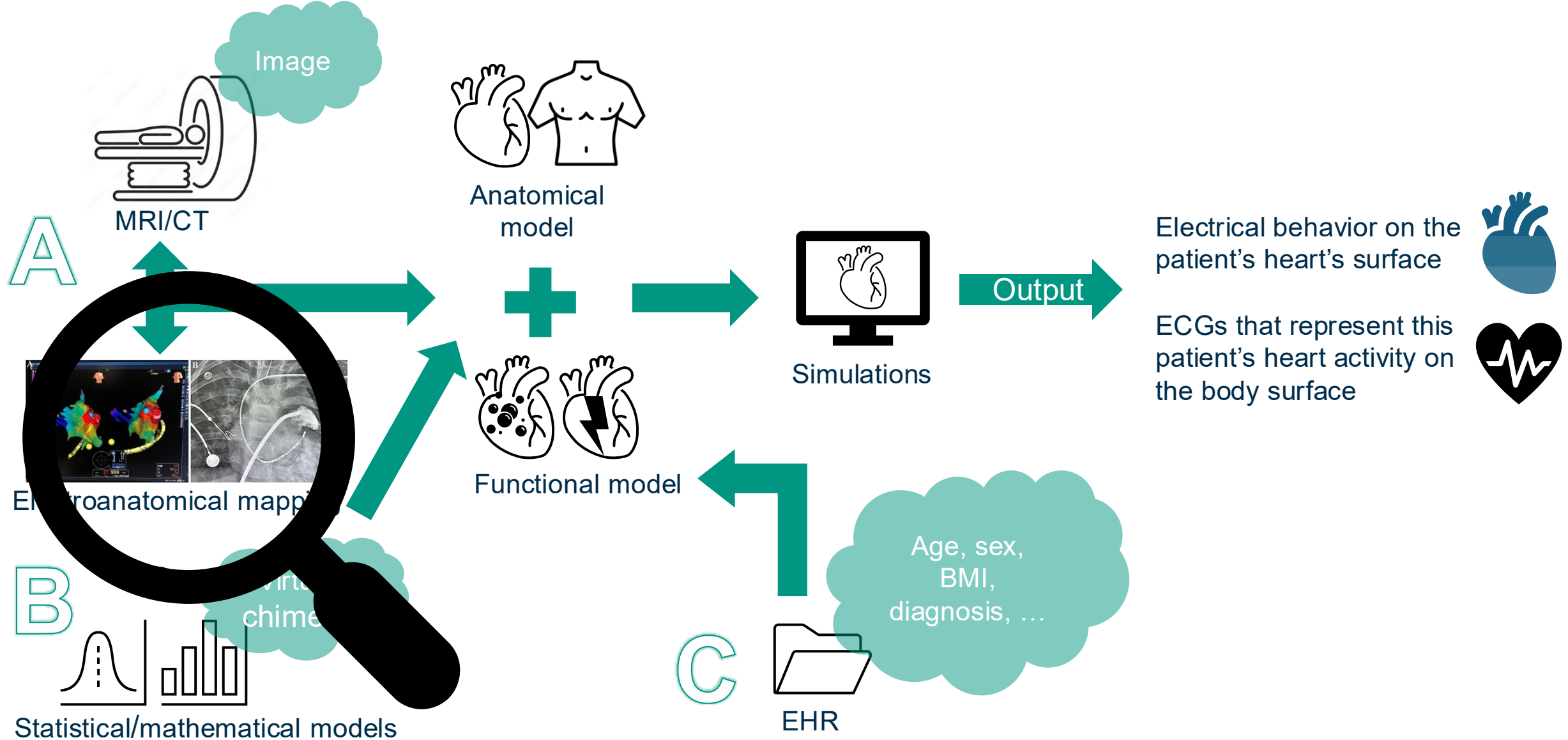
Azzolin et al., 2023, Computerized Medical Imaging and Graphics



Digital Twins

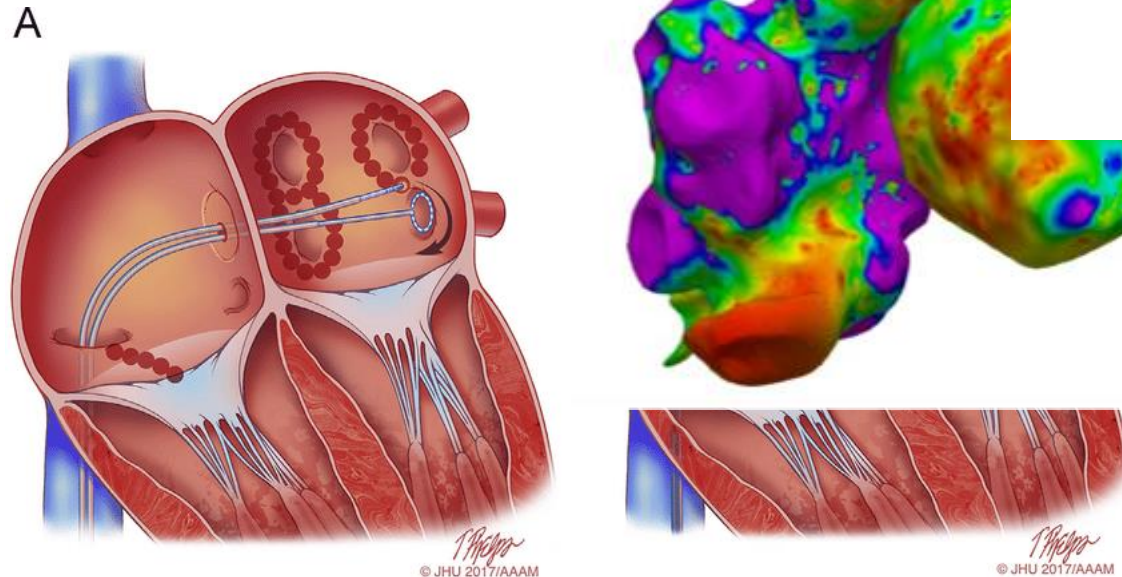


Digital Twins

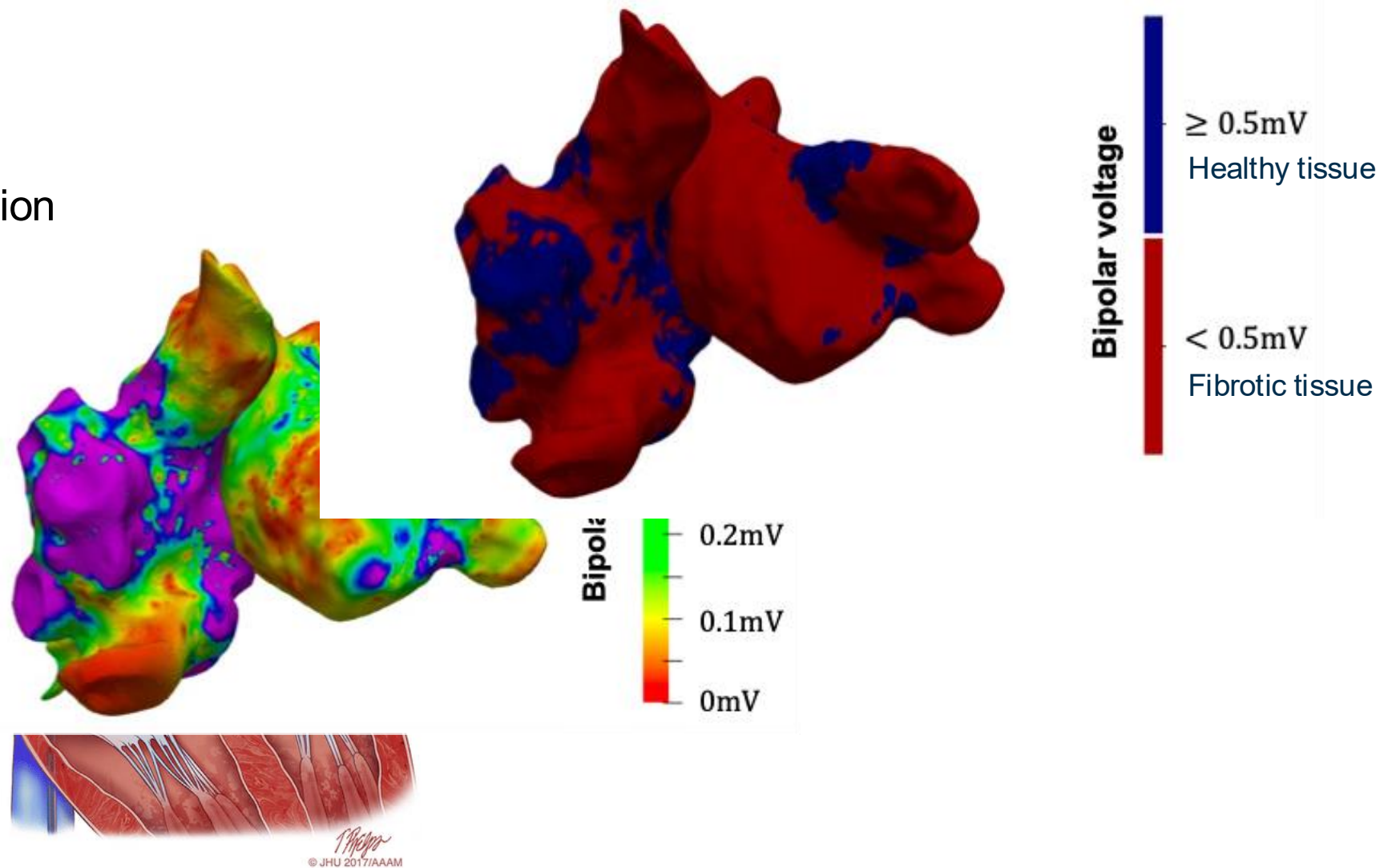


Digital Twins

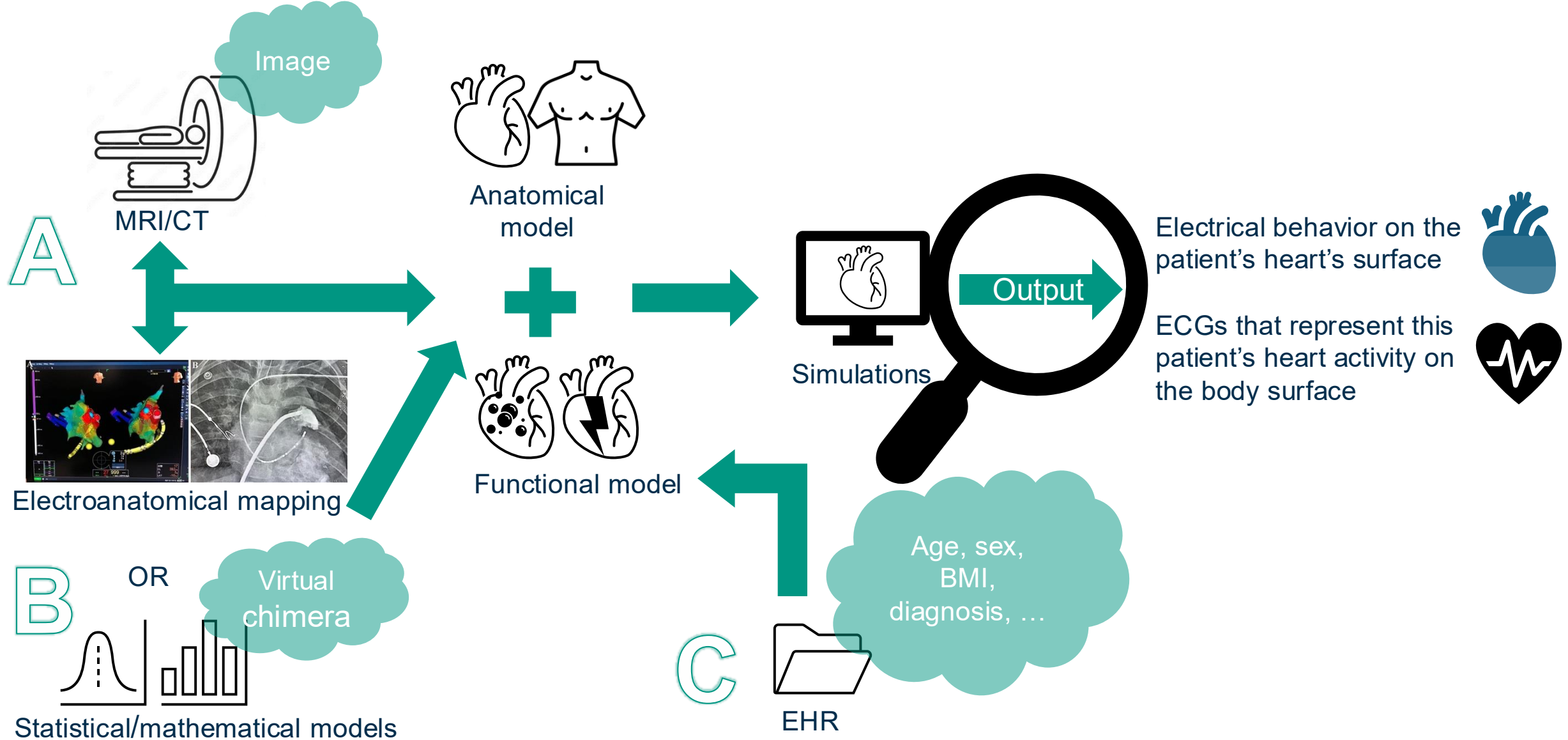
■ Functional personalization



Hopkins, CC BY-NC-ND



Digital Twins



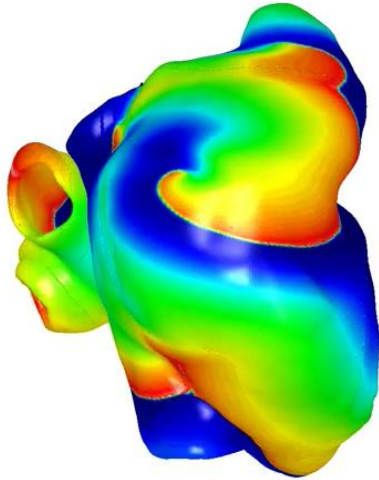
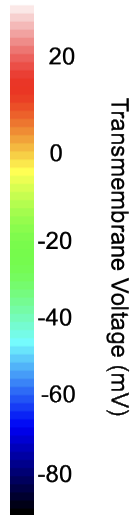
Synthetic ECGs



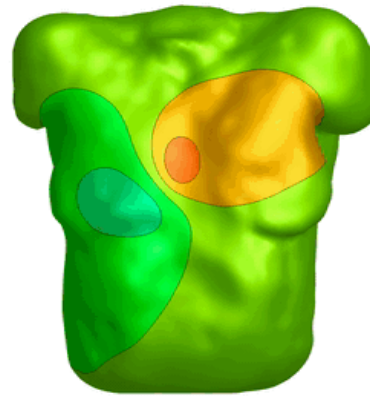
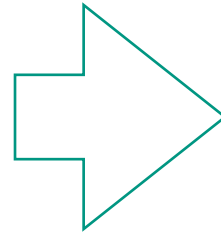
UNIVERSITÄTS
KLINIKUM FREIBURG



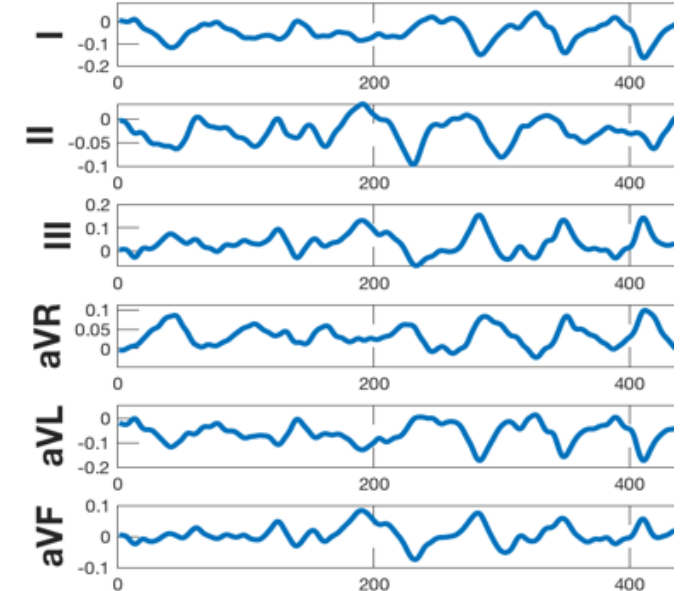
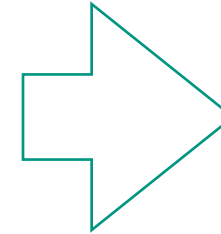
Städtisches Klinikum
Karlsruhe



Atrial fibrillation



Electric potentials
on the body surface



ECG

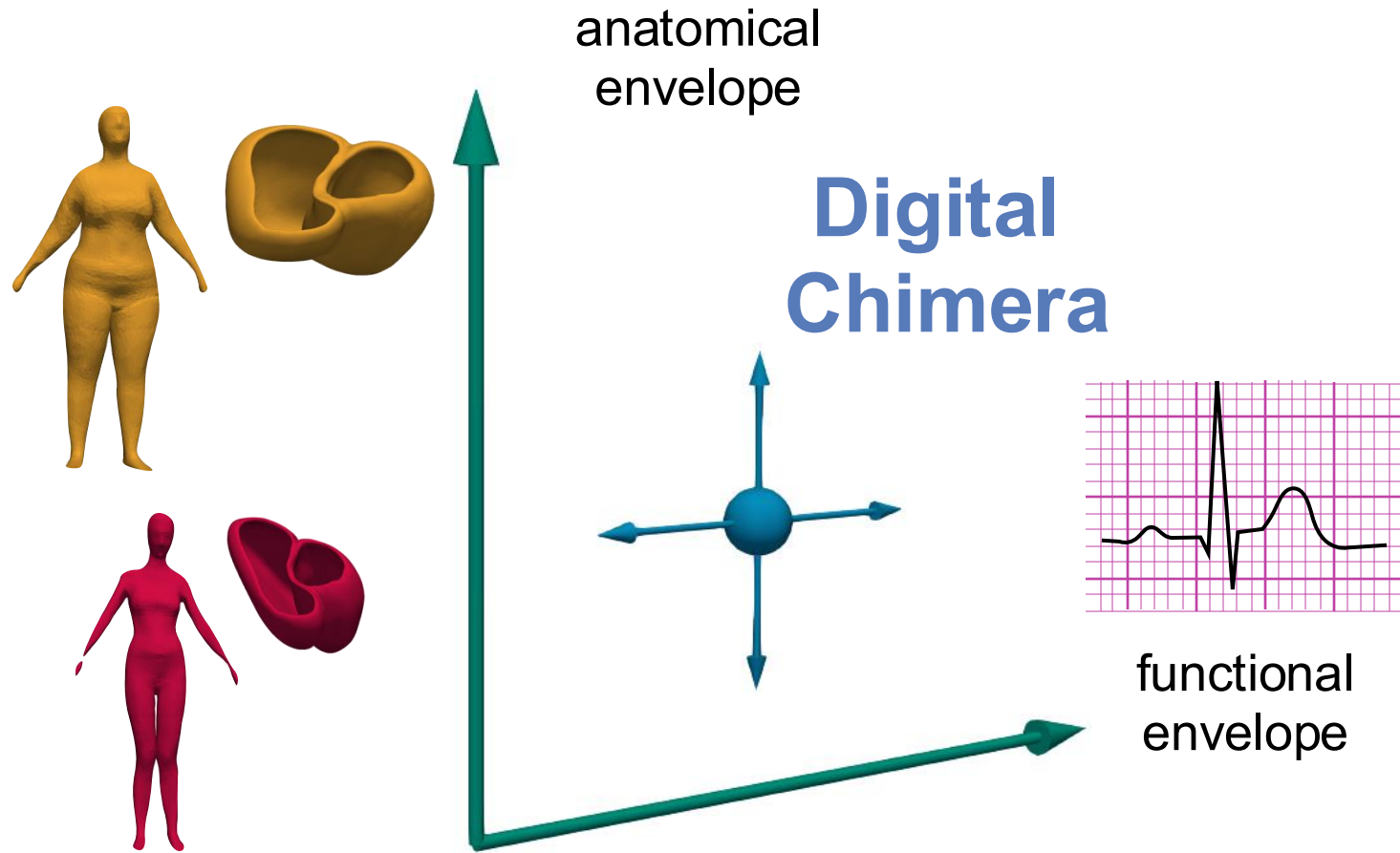


Cardiac Modeling

- Controlled research environment
 - to understand fundamental physiology and pathomechanisms
 - to evaluate diagnostic and therapeutic approaches (in silico studies)
- Scalable research environment to generate big, quality-controlled datasets for machine learning
- Personalized medicine through digital twins



Digital Chimera

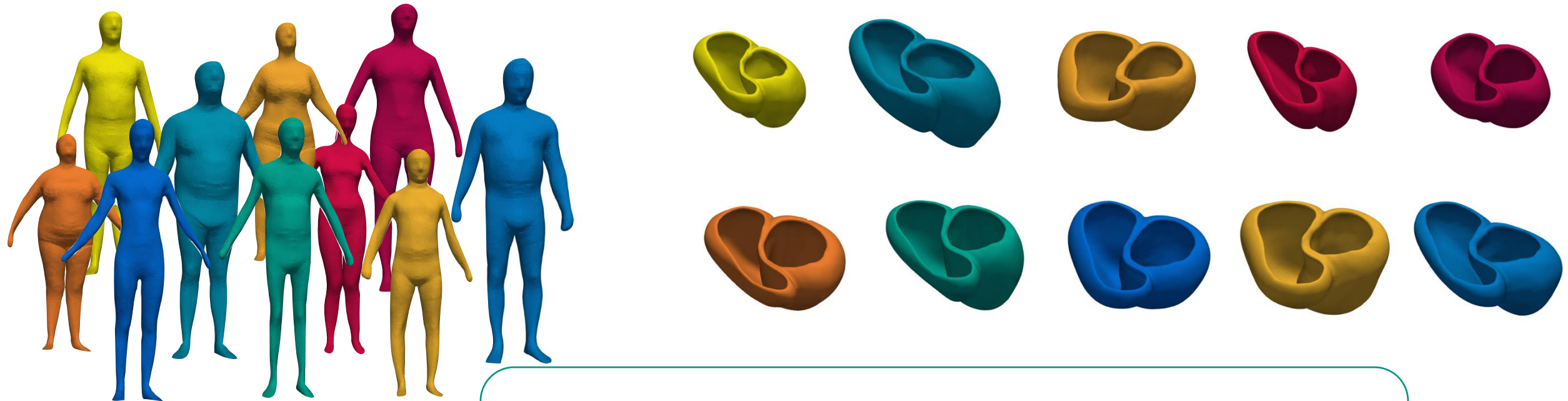


Adapted from Zeike Taylor, University of Leeds



Digital Chimera

■ Anatomical personalization

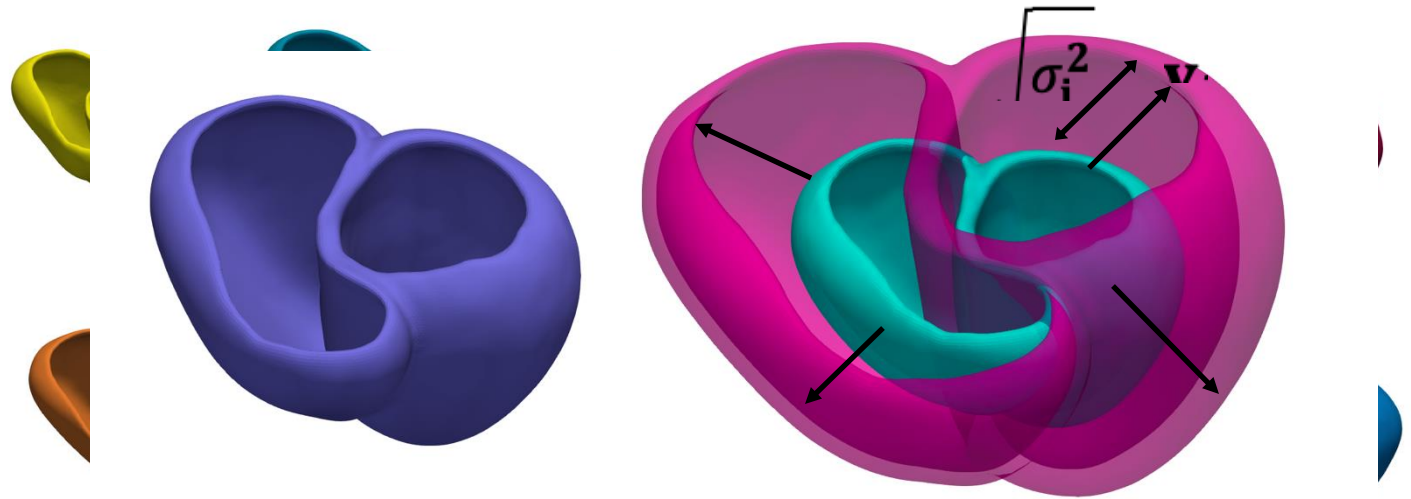
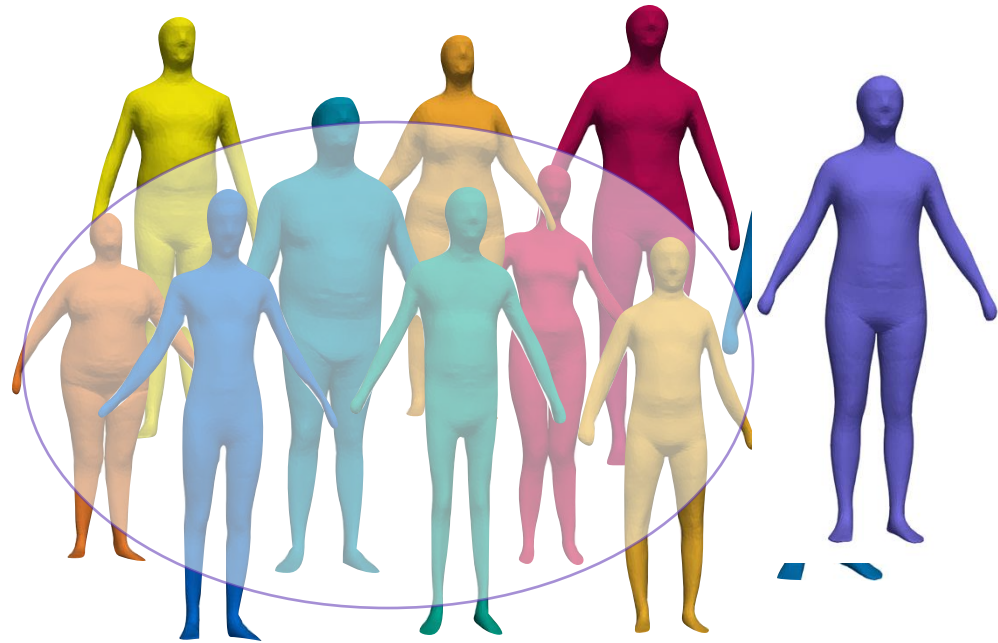


Statistical shape models (SSM) describe the shape variability of an object among a population.



Digital Chimera

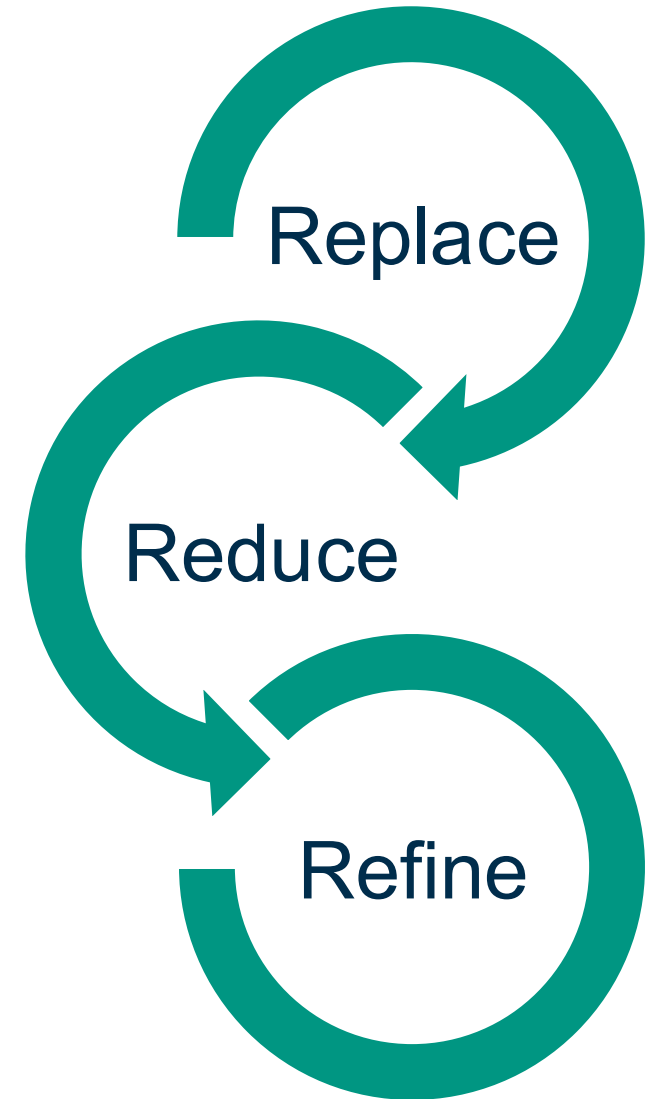
■ Anatomical personalization



$$\mathbf{\Gamma} = \mathbf{\bar{\Gamma}} + \sum_{i=1}^N r_i \cdot \sqrt{\sigma_i^2} \cdot \mathbf{v}_i$$

Potentials of Synthetic Data

- Reduce cost and time required for clinical trials for rare diseases and conditions
 - Contribution to 3Rs of animal testing
- Enhance the predictive power of AI models
- Ensure the delivery of fair treatment recommendations across diverse patient populations
- Enable researchers to access high-quality, representative multimodal datasets without exposing sensitive patient information



Potentials of Synthetic Data

scientific **data**

OPEN

DATA DESCRIPTOR

MedalCare-XL: 16,900 healthy and pathological synthetic 12 lead ECGs from electrophysiological simulations

Karli Gillette^{1,2,8}, Matthias A. F. Gsell^{1,8}, Claudia Nagel^{1,8}, Jule Bender³, Benjamin Winkler⁴, Steven E. Williams^{5,6}, Markus Bär⁴, Tobias Schäffter^{4,5,7}, Olaf Dössel^{1,8}, Gernot Plank^{1,2,8} & Axel Loewe^{3,8} 



Repository KITopen

In silico electrocardiograms of 1.8 million ventricular extrasystoles and corresponding activation maps (part 1)

Pilia, Nicolas ¹; Schuler, Steffen ¹; Rees, Malke ¹; Molk, Gerald ¹; Potyagaylo, Danila ¹; Dössel, Olaf ¹; Loewe, Axel ¹ 

¹ Institut für Biomedizinische Technik (IBT), Karlsruher Institut für Technologie (KIT)

Abstract:

1.8 million ECGs derived from multiscale simulations of cardiac electrophysiology of ventricular extrasystoles. 1000 anatomical variants of a bi-ventricular mesh x 600 excitation origins x 3 heart positions in the torso.

Externe Links

 Download (RADAR4KIT)

Export

 Exportieren als ...

zenodo

Search



Upload

Communities

June 21, 2021

Dataset

Open Access

A Bi-atrial Statistical Shape Model and 100 Volumetric Anatomical Models of the Atria

 Naoel, Claudia; Sanchez, Jorge; Azzolin, Luca; Zheng, Tianbao;  Schuler, Steffen;  Dössel, Olaf;  Loewe, Axel

 Check for updates

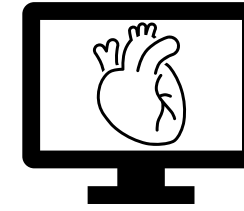


Potentials of Synthetic Data

How can synthetic data enhance the predictive power of AI?

- Simulations can increase dataset size and quality

- Augmentation
- Mix & match
- Ground truth labels
- Balance of classes



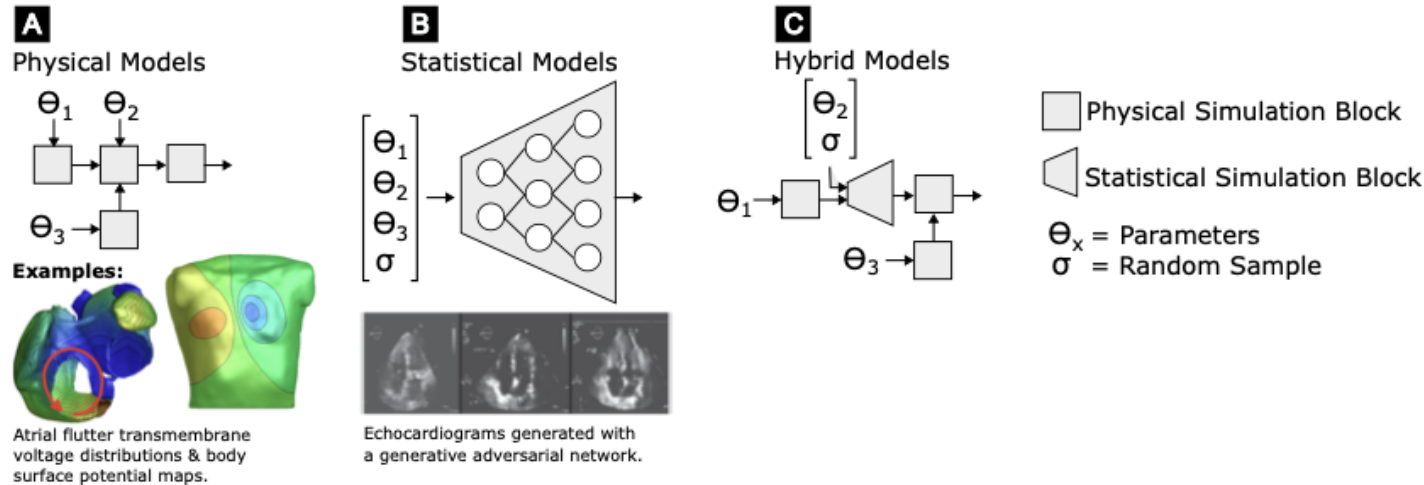
- Computational models

- Controllable
- Quantitative results
- High spatial and temporal resolution
- Mechanistic and physics-based
- Personalizable but not personal per se

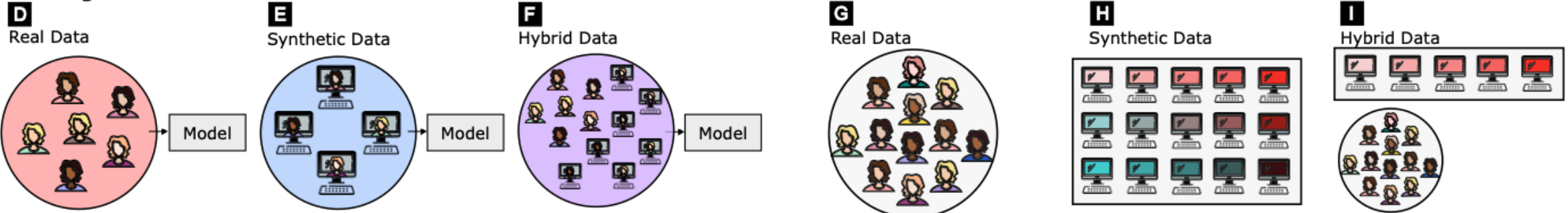


Potentials of Synthetic Data

Simulators



Training



Potential to save rare clinical data for testing



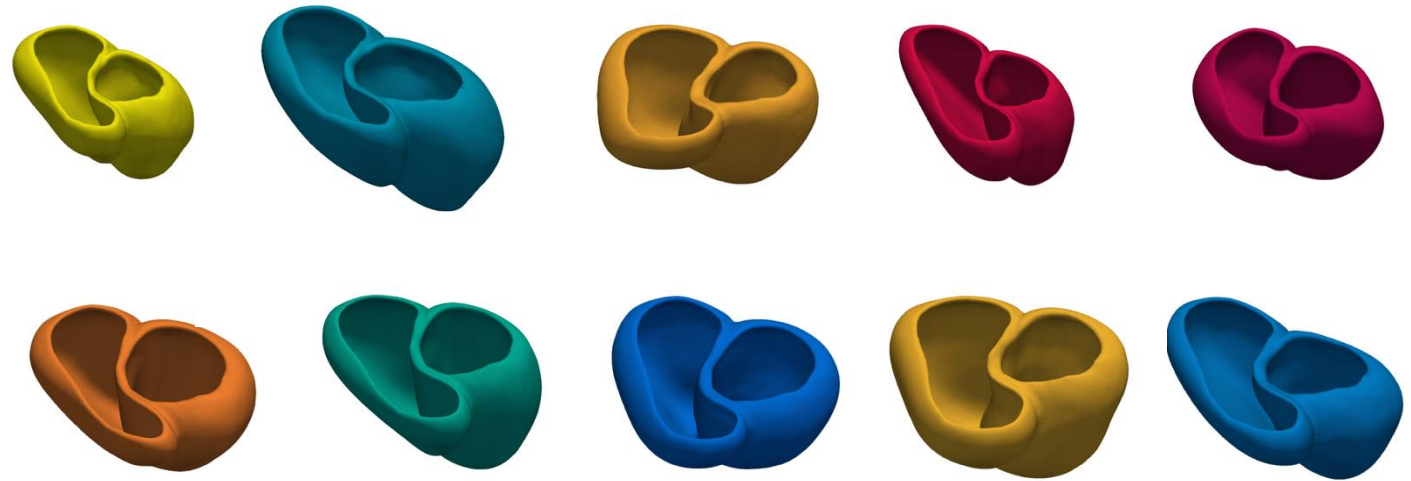
Challenges of Synthetic Data

- Covering relevant variability
- Domain gap (simulation vs reality)
- Computational costs
- Bias is also possible in synthetic data!
- Uncertainty quantification



Clinical Applications

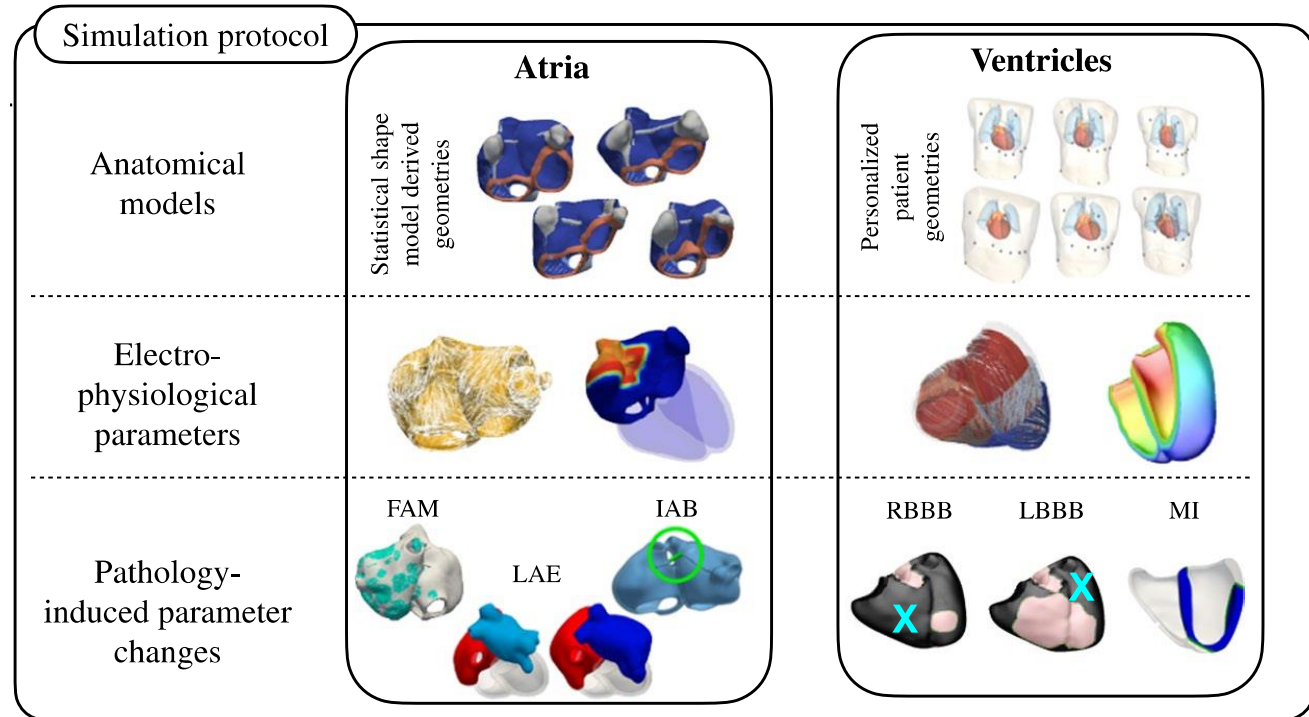
- Covering relevant variability
 - Shape models



Clinical Applications



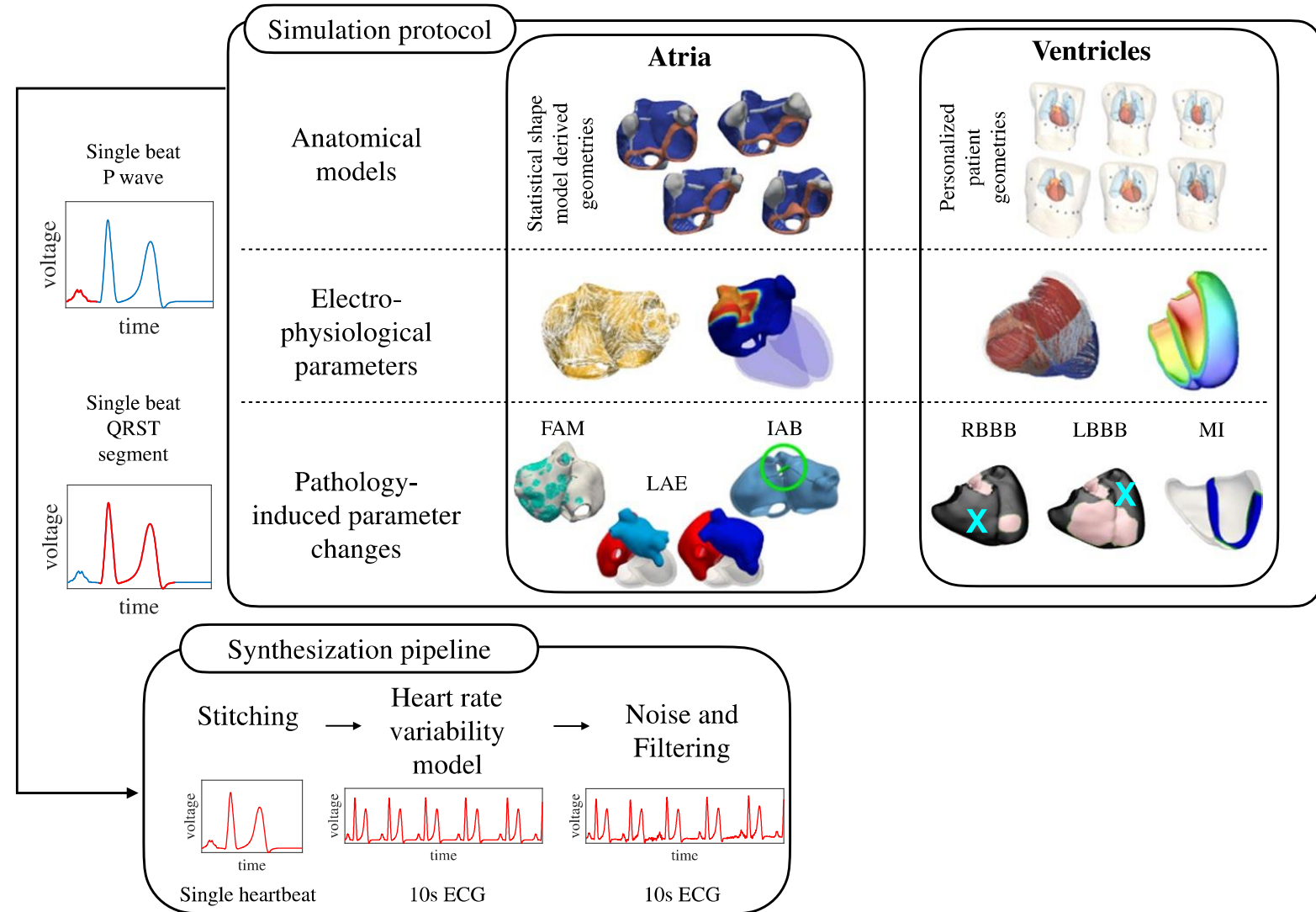
- Closing the domain gap
 - MedalCare-XL



Clinical Applications



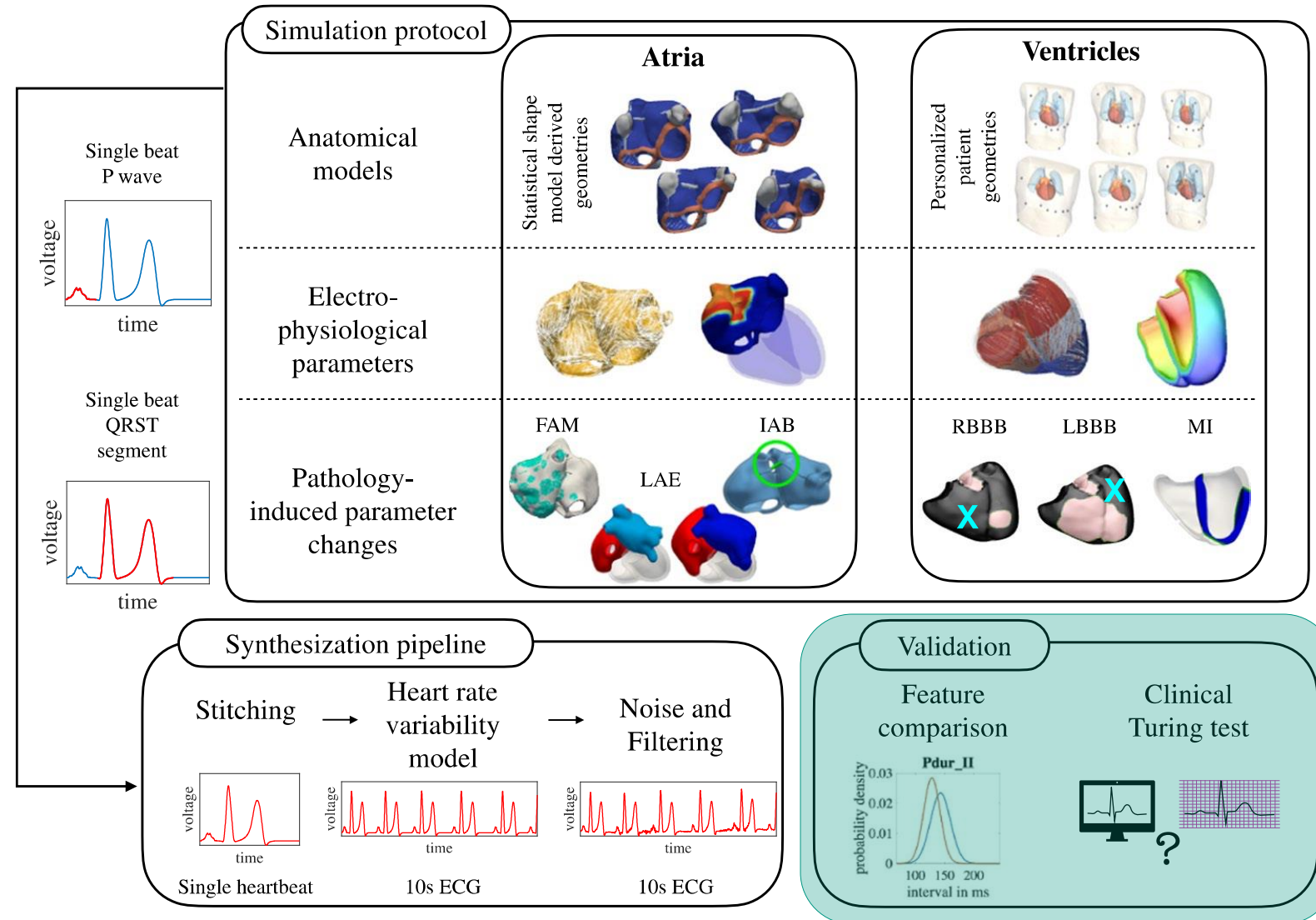
- Closing the domain gap
- MedalCare-XL



Clinical Applications

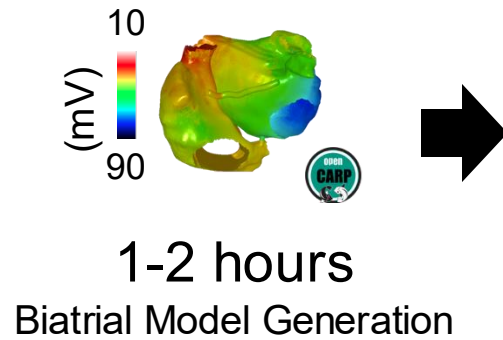


- Closing the domain gap
- MedalCare-XL



Clinical Applications

■ Reducing computational costs



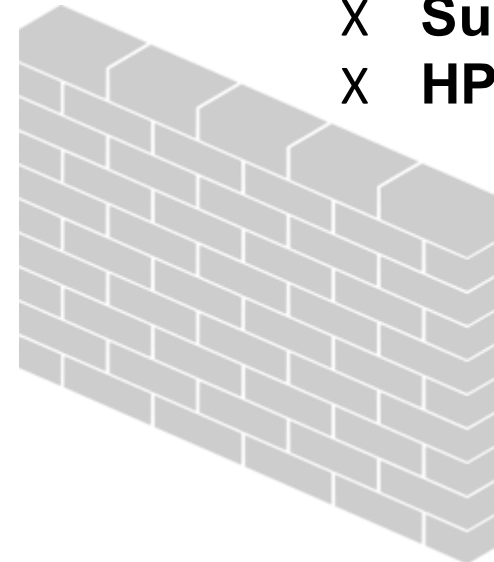
>15 hours
Personalized Digital Twin Assessment

Hospital



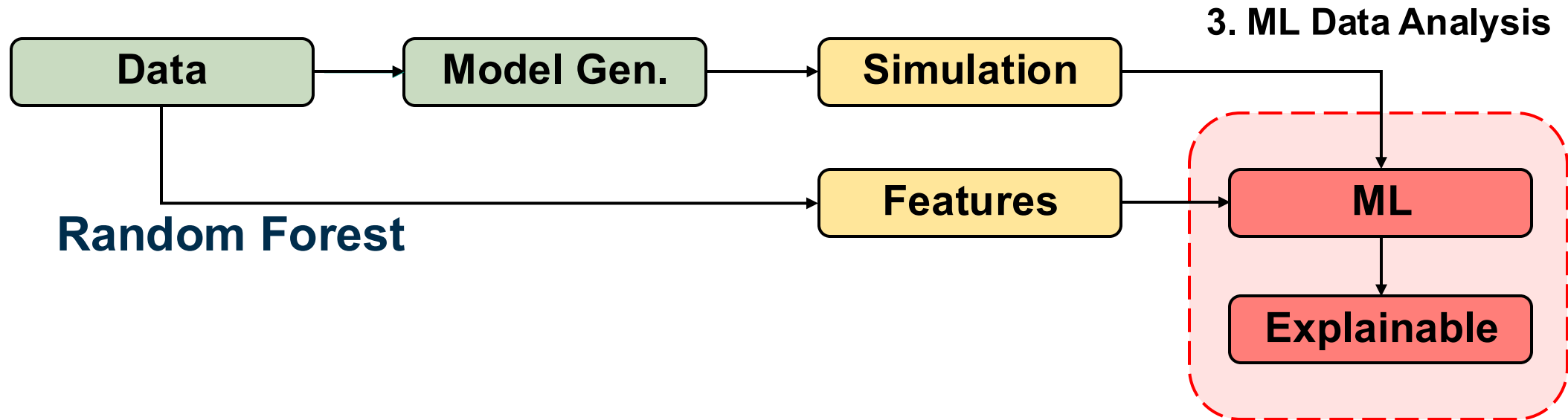
2-3 hours
Electrophysiological Study

X **Substantial time**
X **HPC dependent**



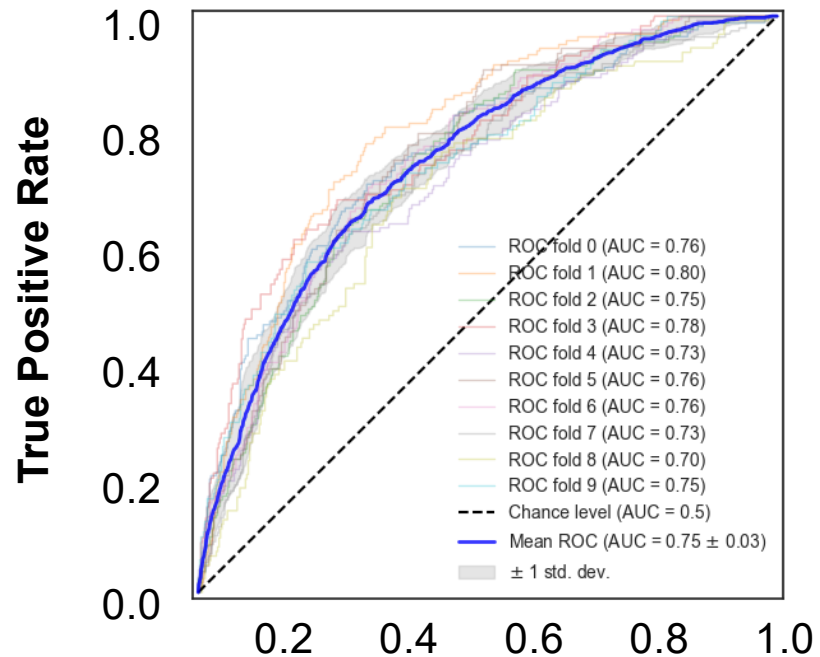
Clinical Applications

- Combination of simulations and machine learning reduces computational costs

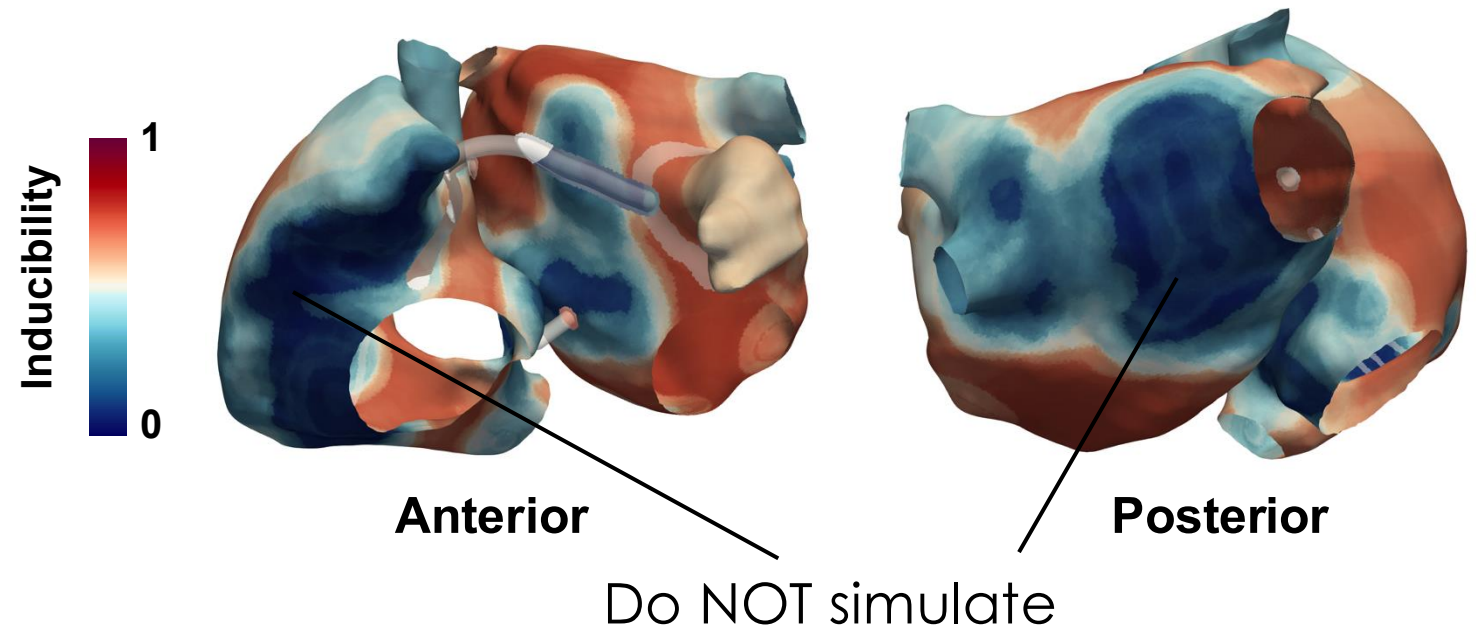


Clinical Applications

Receiver Operating Characteristic (ROC) Curve



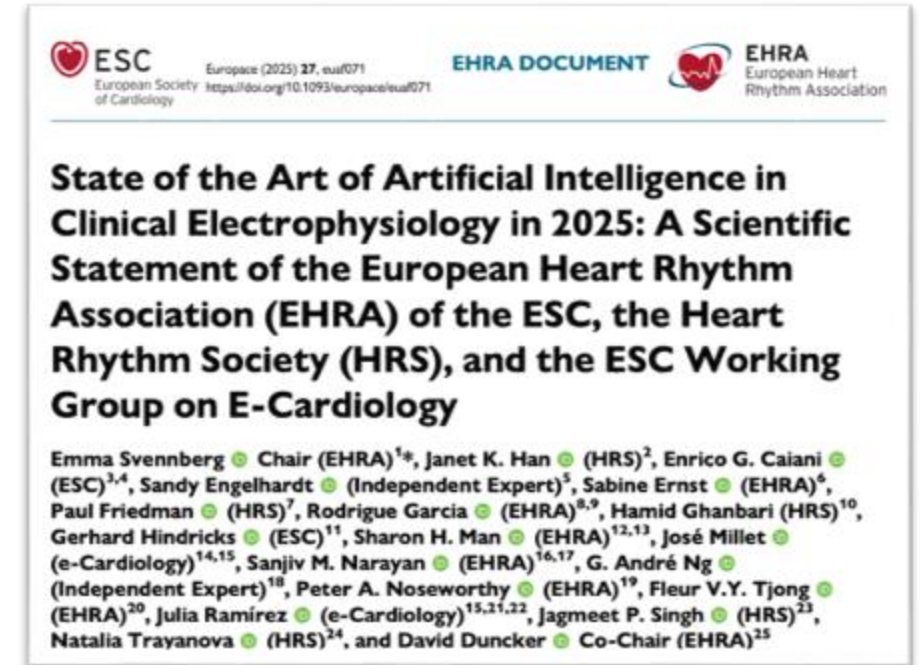
Heat Map for Inducibility Prediction



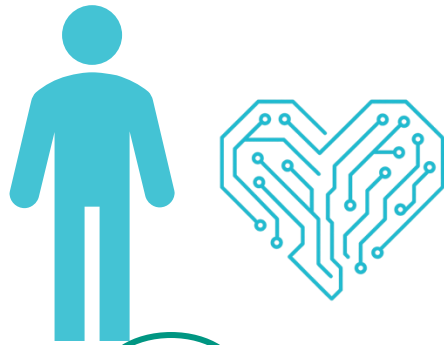
Clinical Applications

EHRA (European Heart Rhythm Association) checklist for AI in clinical electrophysiology

- Data collection and source of data
- Information about training and test data and data split
- Definition of gold standard and ground truth
- Data preparation, data issues
- Balanced groups
- Performance (+compared to classic statistical methods)
- Generalizability
- ...



Opening Example



Patient with rare condition x

Can we also diagnose this person?
How can AI help us?

Rare condition

- ECG features to diagnose might be unclear

➡ Use AI

Few data available

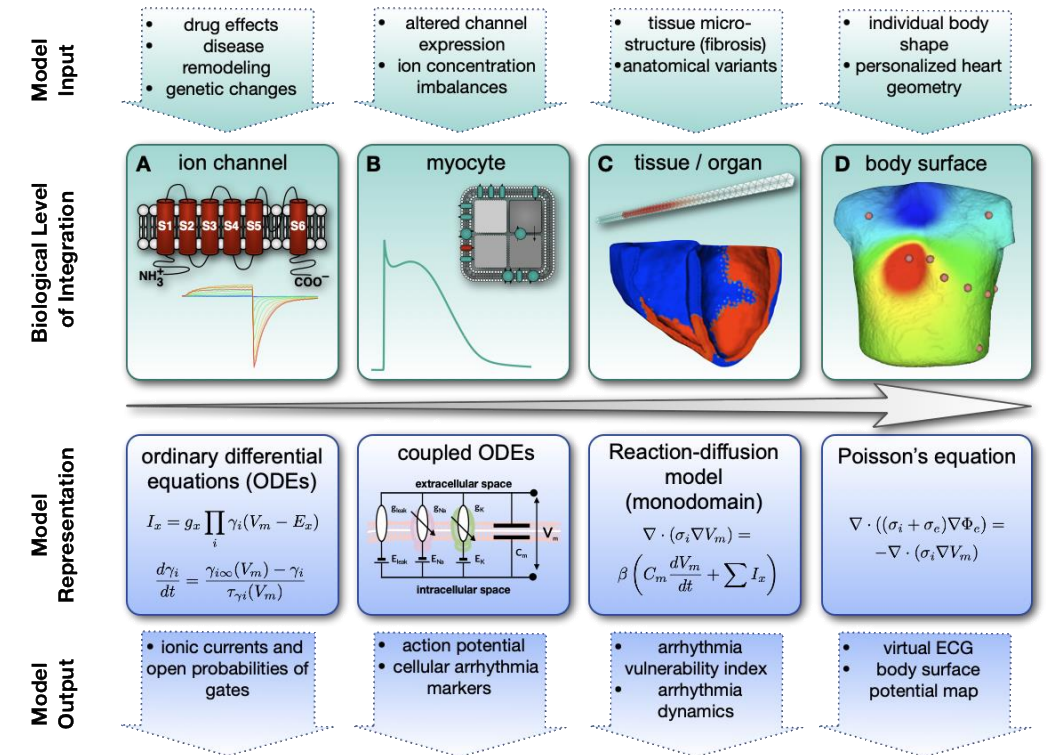
➡ Use synthetic data to increase dataset

- Usage of AI enables screening
 - Objective, scalable, reproducible



Take Home Messages

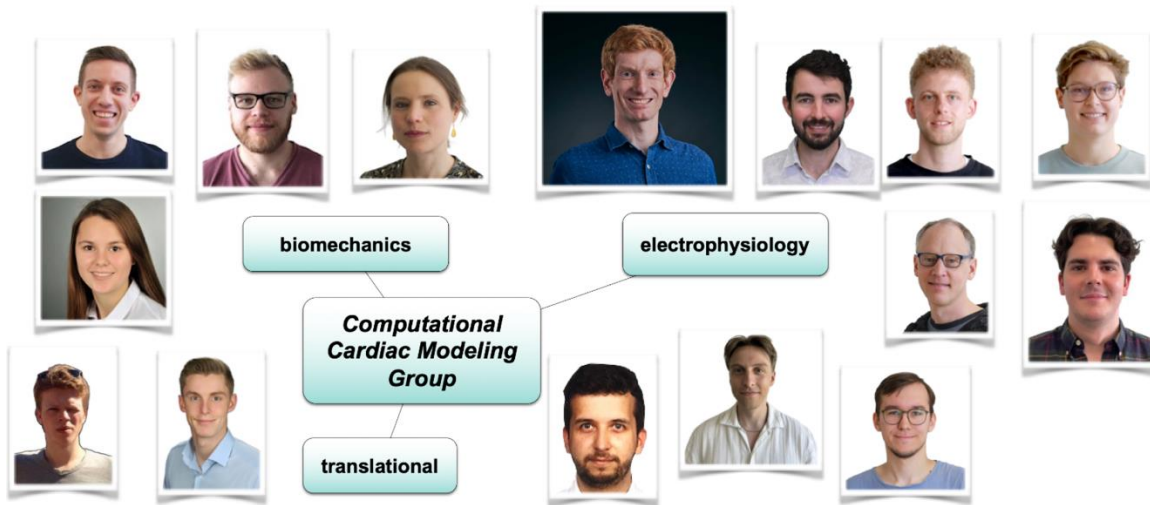
- AI is transforming biomedicine – but data is the fuel
- Synthetic data is a promising tool
 - To increase training data and balance it
 - To understand pathomechanisms
 - To reduce animal experiments & clinical trials
 - To live up to the FAIR principles



- Leibniz ScienceCampus „Digital Transformation of Research“ (DiTraRe)
- Interdisciplinary project with four use cases
- www.ditrare.de
- DiTraRe Symposium on Digitalisation of Research
 - 2-3 December
 - Centre for Art and Media (ZKM) in Karlsruhe
 - Keynote presentations, panel discussions, and interactive sessions



Thank You For Your Attention!



CaMo@KIT

