M2M-InvNet: TMS Induced Electric Field Reconstruction from Muscle Responses Using a Convolutional Network and Variational Inference

Md Navid Akbar, Mathew Yarossi, Sumientra Rampersad, Aria Masoomi, Eugene Tunik, Dana Brooks, and Deniz Erdoğmuş

Abstract— Transcranial magnetic stimulation (TMS) is often applied to the motor cortex to stimulate a collection of motor evoked potentials (MEPs) in groups of peripheral muscles. The causal interface between TMS and MEP is the selective activation of the neurons in the motor cortex; moving the TMS 'spot' around over the motor cortex causes different MEP responses. A question of interest is whether a collection of MEP responses can be used to identify the stimulated locations on the cortex, which could potentially be used to then place the TMS coil to produce chosen sets of MEPs. In this work we leverage our previous report on a 3D convolutional neural network (CNN) architecture that predicted MEPs from the induced electric field, to tackle an inverse imaging task in which we start with the MEPs and estimate the stimulated regions on the motor cortex. We present and evaluate five different inverse imaging CNN architectures, both conventional and variational, in terms of several measures of reconstruction accuracy. We found that one architecture, which we denote as M2M-InvNet, consistently achieved the best performance.

Index Terms— Transcranial magnetic stimulation (TMS), electromyography (EMG), convolutional neural network (CNN), inverse imaging, variational inference (VI).

I. INTRODUCTION

Transcranial magnetic stimulation (TMS) is a non-invasive technique that uses magnetic fields to stimulate neurons in the brain [1]. When TMS is applied to the motor cortex, it may result in muscle activation. This activation can be measured as motor evoked potentials (MEPs) using standard surface electromyography (EMG). By varying coil position over the motor cortex, TMS can be used non-invasively in humans as a causal probe to investigate the spatial topography of muscle activation patterns [2]. TMS mapping of cortical muscle topography has shown clinical utility [3], for example, to quantify cortical muscle topography associated with abnormal muscle activation patterns due to stroke and track changes during recovery [4], [5], and to perform the presurgical evaluation of motor, speech, or language functions for patients requiring resections in eloquent areas [6], [7]. Advances in modeling of the TMS-induced E-field [8], [9] have allowed greater resolution in the estimation of the cortical representations underlying evoked muscle activation. Recently, work from our group has proposed that TMS may be used to study patterns of multi-muscle activation that have been theorized to form the basis of modular control of coordinated movement [10], [11].

Previously, we developed a forward model using a convolutional neural network (CNN) autoencoder (AE) and a separate deep CNN mapper that connects the simulated E-field and recorded MEPs to estimate multi-muscle activation patterns induced by new TMS stimulations [10], [11]. To our knowledge, this was the first report of a robust computational forward modeling framework going from TMS-induced E-Fields to multi-muscle MEPs. In the present study, we expand on our previous forward modeling technique by developing an inverse modeling approach to estimate (putatively causal) cortical E-fields from muscle activation patterns recorded from a collection of relevant muscles. In other words, our system can predict which region of the motor cortex was stimulated by the TMS coil based on a multi-muscle MEP pattern.

We start with subject-specific volume conduction models based on magnetic resonance images (MRIs), followed by finite element (FE) modeling of the E-fields based on the position and orientation of the TMS coil. We report on five deep network architectures that were developed based on selected combinations of CNNs and variational inference (VI). We chose these tools because CNNs have previously been used for TMS modeling to generate head models [12], and to estimate induced E-fields directly from MRI scans [13]. In addition, CNN AEs using VI, known as variational autoencoders (VAEs), have been widely used in computer vision for natural-looking image reconstruction, since deep generative models such as a VAE can constrain the reconstructed image to remain on a learned underlying manifold, such that the reconstructions are more physically or biologically meaningful [14]. VAEs have also matched the performance of standard compressed sensing techniques in inverse imaging with less training data [15]. Three of the five models we developed utilized a two-stage training strategy [15]: first learning a latent space from the E-fields, and second refining that space by learning from the MEP mapping. The remaining two models jointly learned the latent space from the MEPs and the E-fields in a single-stage training strategy [16].

To carry out our study, we collected MRI scans, TMS
coherence, and orientation, and 15-muscle MEP data from
three healthy subjects during expert user-guided cortical motor
mapping. We stimulated at ~1,000 scalp locations
per subject (699, 1200, and 1199 for subjects 1, 2 and 3,
respectively). We used a stratified train-validate-test cross-
validation approach to evaluate the ability of each of these
five networks to accurately estimate the stimulated cortical
region, as determined by the FE modeling, that produced a
given MEP pattern.

Our results suggest that our networks can indeed perform
this task with reasonable accuracy and robustness as long as
there is sufficient MEP activity. The model that directly learns
from cortical stimulation and MEPs jointly achieved the lowest
squared error and the highest fidelity to reconstruction, across
all subjects.

II. METHODOLOGY

All protocols were conducted in conformance with the
Declaration of Helsinki and were approved by the Institutional
Review Board of Northeastern University (IRB# 15-10-22,
last approved September 23, 2021). Three healthy subjects
(3 males, ages 25, 35, & 36) participated after providing insti-
tutionally approved written informed consent. All subjects were
right-hand dominant according to the Edinburgh handedness
inventory [17], free of neurological or orthopedic conditions
that could interfere with the experiment, and met inclusion and
exclusion criteria to receive TMS [18].

A. Data Acquisition

The procedure used for TMS mapping has been previously
described in detail [10]. Briefly, subjects were seated comfort-
ably with the right upper limb supported in an arm trough,
and the left upper limb resting comfortably on an arnrest.
Surface EMG (Delsys Inc., Natick, MA) was recorded at
2000 Hz (common mode rejection ratio >80 dB, 99.99% Ag,
built-in 20–450 Hz bandpass filter) from 15 hand and arm
muscles: 1st dorsal interosseus (FDI), 3rd dorsal interosseous
(3DI), 3rd lumbrical (3Lum), extensor indicus (EI), abduc-
tor pollicis brevis (AbPB), adductor pollicis brevis (AdPB),
abductor digitii minimi (ADM), flexor digitii minimi (FDM),
flexor carpi radialis (FCR), flexor carpi ulnaris (FCU), flexor
digitorum superficialis (FDS), extensor digitorum (EDC), and
extensor carpi radialis (ECR), extensor carpi ulnaris (ECU),
brachioradialis (BRD).

To ensure spatial TMS precision, frameless neuronavigation
(Brainsight, Rogue Research) was used to co-register each
subject’s head position with a 3D cortical surface rendering
of their high-resolution anatomical MRI scan (T1-weighted,
T1 = 1100 ms, TE = 2.63 ms, TR = 2000 ms, 256×192×160
acquisition matrix, 1 mm³ voxels). The TMS coil (Magstim
Rapid2, D70² 70 mm figure-of-eight coil) was held tangential
to the scalp with the handle posterior 45° off the sagi-
tal plane [19]. Motor evoked potentials were measured as
the peak-to-peak EMG amplitude 10-50 ms after the TMS
pulse [5], [10], [20]. The FDI muscle hotspot was found via a
coarse map of the hand knob area to identify the location that
produced the largest and most consistent MEP amplitudes [21].

Resting motor threshold (RMT) was selected as the minimum
intensity required to elicit MEPs >50 µV on 3 out of 6
consecutive stimulations [5]. TMS maps were collected at
stimulus intensities of 110%, 120%, 130%, and 140% of RMT.
The distribution of both the number of stimulations chosen
after preprocessing and those originally applied, corresponding
to each map for each subject, is reported in Table I. The details
of these preprocessing techniques are outlined in Section II-
C. For each map, TMS (100-300 stimulations, 4 s ISI) was
delivered over a 6x6 cm regular grid (1 cm spacing, 36 cm²
area) centered on the hotspot. For each intensity, one stimulus
was delivered to each of the 49 equidistant points on the grid,
and the remaining stimulations were delivered using real-time
feedback from the MEPs to maximize information about the
responsive areas [5], [22], [23]. Care was taken to ensure that
the mapping included the full extent of the excitable area for
all recorded muscles. For each pulse, MEP amplitudes were
recorded of the 15 muscles selected for analysis.

B. Finite Element Modeling of TMS E-fields

The model for subject 1 was constructed as described
in [10] and the simulation data from that study was used for
the analyses in this paper. The models for subjects 2
and 3 were constructed specifically for this work. For these
models, individual T1 MRI scans (TR = 1.9 ms, TE = 2.0 ms,
256×256×176 mm, 1 mm³ voxels) were processed by the
SimNIBS headreco algorithm [24] to produce tissue segmen-
tations of skin, bone, skull cavities, eyes, cerebrospinal fluid,
grey and white matter. Segmentation masks were manually
corrected in Corview (MARREK Inc., Salt Lake City, UT),
converted to surface meshes by headreco, and combined into a
tetrahedral volume mesh with TetGen [25], resulting in meshes
with 3.6 and 3.8 million elements. Isotropic conductivity
values were assigned to each element based on tissue type [26].
The TMS coil was modeled in SCIRun [27] using the Brain-
Stimulator toolbox [28] by approximating the coil field as
the magnetic vector potential from small magnetic dipoles
distributed across the coil [29]. Simulation of electric fields
induced by this coil in the head model was conducted using
a quasi-static FE framework implemented in BrainStimulator.
The resulting E-field for each coil position was then spatially
resampled to a 64×64×64 hexahedral mesh (1 mm³ elements
surrounding the BA4 area of the motor cortex, which was
identified using Freesurfer [30]. This output was used as input
for the next stage of the modeling procedure.

<table>
<thead>
<tr>
<th>Subject</th>
<th>% of RMT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>110</td>
</tr>
<tr>
<td>1</td>
<td>396/300</td>
</tr>
<tr>
<td>2</td>
<td>298/300</td>
</tr>
<tr>
<td>3</td>
<td>299/300</td>
</tr>
</tbody>
</table>

* The numbers are represented in a p/q format, where q indicates the total number of stimulations applied and p indicates the actual number chosen after preprocessing.
C. Preprocessing

Several steps were taken to clean and prepare the data. First, some of the stimulations did not produce any muscle activation. These stimulations were identified automatically (by checking if all the normalized MEP values were zero) and subsequently removed from further analysis to avoid null space problems for the network. Some outlier stimulations that resulted in unusually low E-field values in the BA4 area (maximum simulated E-field intensity <10 mV/m), which we believe were due to experimental errors, were also removed. These stimulations were only applicable for subject 1, and constituted of 49 out of the 149 total for 120% of RMT. Eventually, the actual number of stimulations chosen after applying these preprocessing steps are also reported in Table I.

For network training and testing, min-max scaling was used to preprocess the data. For the E-fields, the intensity of the voxel corresponding to the maximum strength of the E-fields in the entire set of stimulations for a particular subject was scaled to 1, voxels outside the BA4 motor cortex area were scaled to zero, and all other voxels were linearly scaled in that range. For the MEPs, each individual muscle activation was scaled to the unit interval [0,1], with 1 representing the maximum activity of that muscle in the entire set of stimulations, for a given subject.

D. Latent Variable Model

The causal forward model (E-field to MEP mapping) in [11] was expressed as

\[ \tilde{y}_i = f_D(X_i) + \epsilon_{D,i}, \]  

where \( y_i \) is the \( m \times 1 \) observed muscle activity vector for the \( i \)-th stimulation, \( X_i \in \mathbb{R}^{l_x \times l_y \times l_z} \) is the 3D E-field distribution on the motor cortex, \( f_D(\cdot) \) represents the CNN forward model (M2M-Net) [11] for direct mapping of cortical E-fields to MEPs, \( \epsilon_{D,i} \) is the \( m \times 1 \) residual mapping error and \( i \in \{1, 2, ..., I\} \) represents the index of the train or test stimulation. In this work, \( m = 15 \) and \( l_x = l_y = l_z = 64 \). Since the objective of this work was to reuse a similar architecture as M2M-Net to reconstruct the E-field from the MEPs along an inverse imaging path, we sought to obtain an \( f_D^{-1}(\cdot) \) model such that

\[ f_D^{-1}(\tilde{y}_i) = \tilde{X}_i, \]

\[ \approx \tilde{X}_i + f_D^{-1}(\epsilon_{D,i}), \]

where \( \tilde{X}_i \) is the predicted E-field distribution.

Fig. 1 outlines the relations among the different variables, in the causal forward path and inverse imaging path. Here, \( X = X_{1:i} \) are the volumetric E-fields, \( y = \tilde{y}_{1:i} \) are the corresponding muscle activity vectors, and \( z \) is a \( n \times 1 \) latent variable vector, which is assumed to represent the individual subject’s cortico-motor mapping. The inverse imaging model \( f_D^{-1}(\cdot) \), M2M-InvNet, consisted of a mapper block and a decoder block (both with and without an accompanying encoder). While in [11] only a standard AE was explored, four additional CNN architectures with and without VI were tested in this work, based on the idea that deep generative models (such as a VAE) might constrain \( z \) to remain on a learned manifold such that the reconstructions are more accurate [14].

In this work, we designed five inverse models, namely: (a) AE-Decoder, (b) Direct Convolutional, (c) VAE-Decoder, (d) VAE-Sampler-Decoder, and (e) Direct Variational. Models (a), (c) and (d) utilize a two-stage training: they are first trained along a forward path (\( X \rightarrow z \rightarrow y \)), and then the learned \( z \) is utilized to guide the training in the reverse path (\( y \rightarrow z \rightarrow X \)), as seen in Fig. 1. In the forward path, model (a) uses an AE whereas models (c) and (d) use a VAE. Model (c) uses variational sampling only in the forward path, whereas model (d) loads the saved variational sampling from the forward path and re-trains it in the reverse path. Models (b) and (e) implement a single-stage training (\( y \rightarrow z \rightarrow X \)), with (b) using a purely convolutional architecture and (e) using a variational convolutional architecture.

Starting with the forward path in Fig. 1, the goal was to maximize the density function \( P(X) \) from a conditional distribution \( P(X|z) \) [31] as

\[ P(X) = \int P(X|z)P(z)dz, \]

where \( z \) is to be sampled from the density function \( P(z) \). In a standard AE, \( P(z) \) is estimated by the encoder as \( P(z|X) \), whereas \( P(X|z) \) is approximated by the decoder. In a standard VAE, a surrogate distribution \( Q(z|X) \) is used to approximate \( P(z|X) \). To minimize the distance between \( Q(z|X) \) and \( P(z|X) \), the Kullback–Leibler (KL) divergence between them is minimized in a standard VAE [31] as

\[ \min_z KL(Q(z|X)||P(z|X)) = \min_z KL(Q(z|X)||P(z)) \]

\[ -E_{z~Q}[\log P(X|z)]. \]

Moving to the inverse path in Fig. 1, the goal was to map \( z \) from \( y \). The AE-Decoder (a) mapper estimated \( P(z|y) \), whereas the VAE-Decoder (c) and the VAE-Sampler-Decoder (d) mappers estimated \( Q(z|y) \). Subsequently, these three models utilized the saved \( P(X|z) \) decoder, from the forward training, to complete their reverse training. The Direct Convolutional model (b) trained from \( y \rightarrow z \rightarrow X \) directly in a single step, without using a saved pre-trained \( P(X|z) \) decoder. Finally, since the Direct Variational model (e) also trained in a single step, the optimization objective for the...
model became
\[
\min \text{KL} \left[ Q(z|y) || P(z|X) \right]. \tag{5}
\]
In accordance with the right-hand side of (4), (5) may be rewritten to approximate
\[
\min_{z} \text{KL} \left[ Q(z|y) || P(z) \right] - E_{z \sim Q} \left[ \log P(X|z) \right]. \tag{6}
\]

\section*{E. Model Training and Testing}

Corresponding to the models discussed in Section II-D, the family of deep networks developed and compared for this inverse imaging task were instantiated in terms of forward and inverse training paths, as seen in Fig. 2. The forward training paths for each architecture are indicated in the figure with solid arrows, and bold letters beneath the arrows identify architectures that follow that path, while the reverse training paths are shown with dashed arrows and italicized letters beneath the arrows. A bold italicized letter indicates that both the forward and reverse training paths for the specific model take the same route. Each box represents a particular component of the neural network. The numbers above the arrows represent the dimensions of the variables moving between two blocks.

The forward training paths (a-e) begin with the preprocessing block (lower left) in Fig. 2. Coil parameters are chosen at random from a training set of TMS stimulations and the corresponding E-field distribution in the chosen BA4 area is estimated using the finite element simulation. A subject-specific Brodmann area 4 (BA4) binary motor mask is then applied to this E-field distribution. The resulting simulated E-field inside this mask is used as the input \( \tilde{X} \) to the rest of the training network. For the AE-Decoder model (a), the forward path consists of three convolution and activation layers in both the encoder and the decoder blocks, with two max-pool layers in between the convolutional layers in the encoder and two up-sample layers correspondingly positioned in the decoder. For the Direct Convolutional (b) and Direct Variational models (e), the forward paths directly copy over the input simulations to the reconstruction, skipping the encoder and decoder blocks entirely. For the VAE-Decoder (c) and VAE-Sampler-Decoder models (d), the encoder forward paths consist of an additional flattening and variational sampling and reshaping layer. After training along the forward paths is complete, the weights of the decoders are fixed and are not updated further.

After the forward training concludes, we begin the reverse training. The reverse training paths (a-e) begin with a muscle activation vector \( \tilde{y}_j \), in the mapper block (lower right) of Fig. 2. Model (c) passes through just the fully connected layers in the mapper block, joins the trained decoder, and completes the rest of the path as indicated in the figure. Models (a,b) pass through additional sets of convolution, activation, and max-pool layers in the mapper block, before being fed to the trained decoder. Finally, models (d,e) travel through an additional variational sampling layer, before completing similar paths through the decoder as in the forward training. Once training on the reverse paths was complete, the weights in the mapper blocks were also fixed and the networks were ready for inference.

During training, the inputs were processed in mini-batches of size 8. Adadelta was chosen as the optimizer for all models, with a learning rate of 1 to start. Model weights were saved every epoch until the training loss plateaued. If training loss plateaued for five epochs, a dynamic learning rate scheduler multiplied the learning rate by a factor of 0.7. After 20 epochs of no improvement, by at least a factor of \( 10^{-5} \) in the relevant loss function, the training was stopped. Training, evaluation, and output visualization of all models were done on a workstation equipped with a 9th generation Intel Core-i7 3.6 GHz CPU, 64 GB of RAM, and an NVIDIA RTX 2080 Ti GPU hardware. The programming platform used was Python 3.7.4 within JupyterLab, with support from major libraries such as TensorFlow 2.0.0, scikit-learn 0.23.2, matplotlib 3.3.2, etc. The software used for the visualization of the results was MATLAB R2019b.

During each \( j \)-th stimulation in model testing, a muscle activation vector \( \tilde{y}_j \), is chosen from random from a set of muscle activation vector test samples (a set separate from the samples used in training, details of which are outlined in Section II-F) is fed as an input to the inference path, as seen in Fig. 3. The various models follow the paths indicated by the solid arrows in the figure, using the fully trained mapper and decoder. The output from the inference path then produced an estimate of the three-dimensional E-field reconstruction \( \tilde{X}_j \) corresponding to any MEP test sample.

\section*{F. Model Parameters}

The entire set of input-output data for each subject was divided into train and test sets in a 10-fold outer cross-validation (CV) arrangement. This division was stratified such that the distribution of the stimulations from the different levels of stimulation intensity (%RMT), as present in the original data set for a given subject, was preserved between individual train and test sets. The number of channels for the convolutional filters and the value of the \( \ell_1 \) regularization parameter was determined by following [11]. For choosing new parameters, such as the length of the variational sampling layer or the final activation function, we took the training data portion (of subject 1 only) of one of the original folds and subdivided into a second 10-fold "train-validate" CV, where 9/10'ths of each fold was used to train and the last 1/10'th was used as validation. The lowest normalized root mean square error (NRMSE) performance across these 10 validation sets determined the best choice for tuning the relevant parameters. Once tuning parameters were fixed, all models were trained on the entire applicable training set, in each CV fold.

Referring to Fig. 2, there are two sets of convolution-activation-maxpool layers followed by a single convolutional layer in the encoder and two sets of convolution-activation-upsample layers followed by a single convolutional layer in the decoder. The number of channels in the first two convolutional layers of the encoder was 32 and 64, respectively, while for the decoder it was 64 and 32, respectively. The first two convolutional layers in both the encoder and the decoder had \( 3 \times 3 \) filters, while the last one had a \( 1 \times 1 \times 1 \) filter and a single channel. The padding used was \( 1 \) element on each side, and the stride was \( 1 \).

Activation functions followed all convolutional and fully connected layers. The rectified linear unit (ReLU) was chosen...
The different models were first trained along the forward paths (red arrows), and then along the reverse paths (blue arrows), as applicable. The forward training starts in the preprocessing block, continues to the encoder block and then ends in the decoder block. The reverse training begins in the mapper block, and then finishes in the decoder block. The numbers above the arrows indicate the dimensions of the variables moving between any two blocks, while the letters (bold or italicized) below refer to the structures of each of the five architectures as explained below. A bold, italicized letter indicates that both the forward and reverse training paths for a model take the same route. A number shown following an asterisk (e.g. *3) indicates the number of times a layer is present inside a particular block. The five architectures are denoted as: (a) AE-Decoder, (b) Direct Convolutional, (c) VAE-Decoder, (d) VAE-Sampler-Decoder, (e) Direct Variational.

Max-pooling and upsampling layers were used to reduce and increase the sizes of the representations, respectively, and had filter windows of size $2 \times 2 \times 2$ and a stride of 1.

The E-field distributions were sparse in the BA4 volume for any given stimulation, and since an $\ell_1$ penalty has been shown to be effective in convolutional sparse coding [33], we added an $\ell_1$ regularization of value $10^{-4}$, as we did earlier in [11].

To train the forward path only for the VAE-Decoder (c) and both the forward and reverse paths for the VAE-Sampler-Decoder (d), the relevant objective was to minimize a combination of the MSE loss and the KL divergence loss from (4), given by

$$L_2(\theta) = \text{MSE}_{\text{loss}} + \text{KL}[Q(z|X)\|P(z)].$$

G. Loss Functions

Each of the five networks was trained to optimize a relevant cost function. To train both the forward and reverse paths for the AE-decoder (a) and Direct Convolutional models (b), as well as the reverse path only for the VAE-decoder (c), we minimized the mean squared error (MSE) loss, which we denote as $L_1(\theta)$, between the ground truth (GT) E-field distribution and its reconstruction, for $N$ training samples as

$$L_1(\theta) = \frac{1}{N} \sum_{i=1}^{N} \| \hat{X}_i - \tilde{X}_i \|^2_2.$$ 

where $\| \cdot \|_2$ denotes the Euclidean norm.

To train the forward path only for the VAE-Decoder (c) and both the forward and reverse paths for the VAE-Sampler-Decoder (d), the relevant objective was to minimize a combination of the MSE loss and the KL divergence loss from (4), given by

$$L_2(\theta) = \text{MSE}_{\text{loss}} + \text{KL}[Q(z|X)\|P(z)].$$

as the activation function for all intermediate layers. The final activation function in the decoder was a bounded ReLU, which implemented the minimum value between 1 and a ReLU output [32]. This choice arose from the need to constrain the output between 0 and 1, to match the min-max scaling applied earlier at the input, and maintain the physiological interpretability of the reconstructed E-fields. Although a sigmoid activation serves the same purpose, the bounded ReLU consistently outperformed the sigmoid in the inner CV experiments we conducted and was thus used in all models in this work.
We assume the latent distribution and the surrogate to be Gaussians, parameterized as \( P(z) = N(0, 1) \) and \( Q(z|X) = N(\mu(X), \Sigma(X)) \) \cite{31}. This loss can then be rewritten as

\[
L_2(\theta) = \text{MSE}_{\text{loss}} + \frac{1}{2} \sum_n \left[ \exp\{\log \Sigma(X)\} + \{\mu(X)\}^2 - 1 - \log \Sigma(X) \right].
\] (9)

Finally, to train both the forward and reverse paths for the Direct Variational model (e), an objective consisting of the MSE loss and the relevant expression for the KL divergence loss from (6) was minimized:

\[
L_3(\theta) = \text{MSE}_{\text{loss}} + \text{KL}[Q(z|y)||P(z)].
\] (10)

Assuming \( Q(z|y) \) to be a Gaussian parameterized by \( N(\mu(y), \Sigma(y)) \), (10) can be reformulated as

\[
L_3(\theta) = \text{MSE}_{\text{loss}} + \frac{1}{2} \sum_n \left[ \exp\{\log \Sigma(y)\} + \{\mu(y)\}^2 - 1 - \log \Sigma(y) \right].
\] (11)

### H. Evaluation Metrics

All five models were first trained on each fold’s training set and then evaluated on the corresponding test set for each CV fold. In each CV round, the model weights were first cleared and then randomly initialized for a new iteration of training. The performance of each model was assessed for each of the three subjects, using three evaluation criteria. NRMSE, the primary metric of performance assessment, was calculated for the \( j \)-th test stimulation as

\[
\text{NRMSE}_j = \sqrt{\frac{||\hat{\mathbf{X}}_j - \bar{\mathbf{X}}_j||^2}{||\bar{\mathbf{X}}_j||^2}},
\] (12)

where \( \hat{\mathbf{X}} \) and \( \mathbf{X} \) are as defined in Eq. 7. To measure the similarity between the individual reconstructions of the E-field and the respective GT, \( R^2 \) was calculated as a secondary metric:

\[
R^2_j = 1 - \frac{||\hat{\mathbf{X}}_j - \bar{\mathbf{X}}_j||^2}{||\bar{\mathbf{X}}_j - \bar{\mathbf{X}}||^2},
\] (13)

where \( \bar{\mathbf{X}}_j \) is the mean of all voxel intensities \( (v_{jk}) \) contained in \( \hat{\mathbf{X}}_j \) for voxels \( k \), within the volume of the motor cortex \( K \). Finally, the center of gravity (CoG), a common outcome used in TMS mapping \cite{2}, for both the GTs and the predictions of the E-field distributions were calculated as

\[
\text{CoG}_j = \left( \frac{\sum_k v_{jk}x_k}{\sum_k v_{jk}}, \frac{\sum_k v_{jk}y_k}{\sum_k v_{jk}}, \frac{\sum_k v_{jk}z_k}{\sum_k v_{jk}} \right);
\] (14)

where \( x_k, y_k, \) and \( z_k \) are the Cartesian coordinates \( \forall k \in K \). The error in CoG (\( \text{CoG}_{\text{error}} \)) in the reconstructions then formed the tertiary metric:

\[
\text{CoG}_{\text{error},j} = ||\text{CoG}_{\text{GT},j} - \text{CoG}_j||_2,
\] (15)

where \( \text{CoG}_{\text{GT},j} \) is the GT CoG for \( \hat{\mathbf{X}}_j \) and \( \text{CoG}_j \) is the CoG calculated for \( \bar{\mathbf{X}}_j \).

### III. RESULTS

#### A. Performance Across Models

Table II reports our statistics from the performance comparison of the presented models, across ten cross-validation folds, for all three subjects. The mean NRMSE and \( R^2 \) are reported, across all stimulations for each subject, along with the corresponding standard errors of the mean (SEMs) calculated for a 95% level of confidence. For the first two purely convolutional models, we observed that the Direct Convolutional (a) consistently outperformed the AE-Decoder (b) across all subjects. For the next three models (c,d,e) involving VI, it was noticeable that all of them performed better than their purely convolutional counterparts. Finally, the Direct Variational model (e) consistently performed the best, both in terms of NRMSE and \( R^2 \), with the VAE-Sampler-Decoder (d) as a close second.

To obtain qualitative insight into the E-field reconstruction fidelity, the performance of the various models is illustrated in Fig. 4 for a single stimulation for subject 3 that elicited large responses in most of the muscles. The image on the top left of Fig. 4 shows the normalized GT E-field of the chosen stimulation, with muscle activation vector (normalized as described above) in the inset bar graph. The other five panels show the five different reconstructions. In the reconstructions using the AE-Decoder (a) and the Direct Convolutional (b) architectures (first row), we observe underestimation of the intensity of the E-fields around the CoG, producing flatter intensity profiles than are present in the GT. With the VAE-Decoder (c), where the latent space was learned from the E-fields and subsequently fixed, a similar result was observed.

For the VAE-Sampler-Decoder (d) and Direct Variational (e), we observe that the reconstructions reproduced the GT E-fields with a high degree of fidelity. Although it may seem difficult to distinguish between the two outputs in this specific example, the proposed Direct Variational (e) model consistently outperformed the VAE-Sampler-Decoder (d) model in aggregate, as seen in Table II. Since the Direct Variational model provided the most accurate reconstructions, we present example results using that architecture.

### Table II: Performance of the Different Models, for Each Subject. The Means and Standard Errors of the Means Across the Cross-Validation Folds Are Reported for the NRMSE and \( R^2 \).

<table>
<thead>
<tr>
<th>Subject</th>
<th>Model</th>
<th>NRMSE</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AE-Decoder</td>
<td>0.294±0.014</td>
<td>0.836±0.011</td>
</tr>
<tr>
<td></td>
<td>Direct Convolutional</td>
<td>0.219±0.011</td>
<td>0.832±0.010</td>
</tr>
<tr>
<td></td>
<td>VAE-Decoder</td>
<td>0.208±0.010</td>
<td>0.852±0.011</td>
</tr>
<tr>
<td></td>
<td>VAE-Sampler-Decoder</td>
<td>0.184±0.008</td>
<td>0.854±0.010</td>
</tr>
<tr>
<td></td>
<td>Direct Variational</td>
<td>0.181±0.008</td>
<td>0.855±0.010</td>
</tr>
<tr>
<td>2</td>
<td>AE-Decoder</td>
<td>0.402±0.011</td>
<td>0.692±0.009</td>
</tr>
<tr>
<td></td>
<td>Direct Convolutional</td>
<td>0.252±0.007</td>
<td>0.720±0.009</td>
</tr>
<tr>
<td></td>
<td>VAE-Decoder</td>
<td>0.245±0.006</td>
<td>0.744±0.011</td>
</tr>
<tr>
<td></td>
<td>VAE-Sampler-Decoder</td>
<td>0.220±0.005</td>
<td>0.754±0.010</td>
</tr>
<tr>
<td></td>
<td>Direct Variational</td>
<td>0.213±0.006</td>
<td>0.756±0.010</td>
</tr>
<tr>
<td>3</td>
<td>AE-Decoder</td>
<td>0.250±0.008</td>
<td>0.566±0.012</td>
</tr>
<tr>
<td></td>
<td>Direct Convolutional</td>
<td>0.222±0.007</td>
<td>0.571±0.012</td>
</tr>
<tr>
<td></td>
<td>VAE-Decoder</td>
<td>0.216±0.007</td>
<td>0.595±0.014</td>
</tr>
<tr>
<td></td>
<td>VAE-Sampler-Decoder</td>
<td>0.166±0.005</td>
<td>0.682±0.014</td>
</tr>
<tr>
<td></td>
<td>Direct Variational</td>
<td>0.156±0.005</td>
<td>0.714±0.014</td>
</tr>
</tbody>
</table>

The best value for each metric in each subject is shown in bold.
stimulations with respect to the muscle response profile, we
may explain why the reconstruction performance for this
subject 1 the low E-field stimulations that were discarded
across the stimulation intensities and participants, indicating
subject. We did not observe a clear trend in model performance
averaged across all 10 folds for each intensity and for each

only in the next subsections.

B. Performance of the Direct Variational Model: Effects of Stimulation Intensity

The induced E-field is directly related to the intensity of stimulation. We therefore, analyzed the reconstruction performance of the Direct Variational model with respect to the four stimulation intensities applied. In Table III, we report the mean and SEMs (for a 95% level of confidence) of NRMSE and $R^2$, averaged across all 10 folds for each intensity and for each subject. We did not observe a clear trend in model performance across the stimulation intensities and participants, indicating that Direct Variational model performance, in aggregate, was not sensitive to the stimulation intensity used. We recall that for subject 1 the low E-field stimulations that were discarded in preprocessing, as described in Section II-C, constituted 49 out of the 149 total for intensity of 120% RMT. That may explain why the reconstruction performance for this stimulation intensity did not match that for the other three intensities, in subject 1.

C. Performance of the Direct Variational Model: Effects of Muscle Response

To give insight into differences in performance across stimulations with respect to the muscle response profile, we visualize the best, average, and worst E-field reconstructions, based on NRMSE, for the Direct Variational model in Fig. 5 for the same subject as in the previous figure (subject 3). The best reconstruction, in terms of lowest NRMSE, also yielded a very low CoG$_{error}$ and shift of the CoG (both close to zero) and a very high $R^2$ (close to one). The reconstruction error map for this case confirms that it reproduced the ground truth E-field with very high accuracy, with a maximum normalized voxel intensity of 0.05 in the error map. We note that the mean activation of the input muscles was high across many muscles. In the stimulation with an NRMSE that was closest to the average performance for subject 3 (middle column), as reported in Table II, the NRMSE was higher, the CoG$_{error}$ was larger, and the $R^2$ was smaller than for the best case. This reconstruction also replicated ground truth well, though small artifacts are visible along the edges of the error map (bottom row), with a maximum normalized voxel intensity error of 0.18. The number of activated muscles and the mean activation across muscles was lower in comparison to that for the best case reconstruction. Finally, the worst reconstruction for this subject, shown in the right-hand column, corresponded to a case where the MEP activation was localized to a single muscle with a small amplitude. The NRMSE was substantially higher than for the other two examples, the CoG$_{error}$ was correspondingly large and $R^2$ was low. The error map showed broad regions of high normalized voxel intensities (0.2–0.4) where the errors were high, with 0.43 as the maximum normalized voxel intensity.

To illustrate the effect of the muscle response profile on the performance of the Direct Variational model across test stimulations from all CV folds, we show scatter plots of the NRMSE against the mean (Fig. 6a) and variance (Fig. 6b) of the normalized MEPs for the same subject (subject 3). The highest error samples were largely concentrated where both the mean and variance of activation were the lowest, and NRMSE decreased with increased mean and variance of activation across muscles. As expected, the mean and variance of the muscle response to stimulation increased with increasing intensity, however, the relationship between NRMSE and intensity was variable in agreement with the aggregate data shown in Table III.

To view the effect of the muscle response profile from a different perspective, we show the NRMSE distributions as box-plots (Fig. 6c) against the number of active (non-zero

TABLE III

<table>
<thead>
<tr>
<th>Subject</th>
<th>% of RMT</th>
<th>NRMSE</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>110</td>
<td>0.172±0.015</td>
<td>0.879±0.020</td>
</tr>
<tr>
<td></td>
<td>120</td>
<td>0.213±0.033</td>
<td>0.806±0.052</td>
</tr>
<tr>
<td></td>
<td>130</td>
<td>0.176±0.019</td>
<td>0.863±0.027</td>
</tr>
<tr>
<td></td>
<td>140</td>
<td>0.176±0.021</td>
<td>0.862±0.029</td>
</tr>
<tr>
<td>2</td>
<td>110</td>
<td>0.222±0.015</td>
<td>0.770±0.039</td>
</tr>
<tr>
<td></td>
<td>120</td>
<td>0.204±0.011</td>
<td>0.792±0.035</td>
</tr>
<tr>
<td></td>
<td>130</td>
<td>0.219±0.011</td>
<td>0.743±0.043</td>
</tr>
<tr>
<td></td>
<td>140</td>
<td>0.208±0.011</td>
<td>0.717±0.042</td>
</tr>
<tr>
<td>3</td>
<td>110</td>
<td>0.169±0.014</td>
<td>0.668±0.034</td>
</tr>
<tr>
<td></td>
<td>120</td>
<td>0.152±0.012</td>
<td>0.707±0.032</td>
</tr>
<tr>
<td></td>
<td>130</td>
<td>0.136±0.011</td>
<td>0.756±0.029</td>
</tr>
<tr>
<td></td>
<td>140</td>
<td>0.166±0.013</td>
<td>0.723±0.031</td>
</tr>
</tbody>
</table>
Fig. 5. Comparison of the best, average, and worst reconstructions (in terms of NRMSE) along the columns from left to right, for subject 3, using the Direct Variational model. Starting from the top row, each column shows the normalized input MEPs, the corresponding stimulation intensity, the ground truth (GT) E-field distribution simulated on the brain, the GT E-field (E-field\textsubscript{GT}) on the BA4 map, the reconstructed E-field (E-field\textsubscript{Rec}) on the BA4, and finally an error map (\|E-field\textsubscript{GT} - E-field\textsubscript{Rec}\|) on BA4 to further illustrate the accuracy of the reconstruction. The CoG is indicated on the GT and reconstructed BA4 maps with a red circle. Note that the color map in the simulated E-field is shown in units of V/m, while the intensities of the reconstruction maps are normalized as in the previous figure.

IV. DISCUSSION AND LIMITATIONS

Table II provided evidence that the purely convolutional AE structure [10], [11] and the two-stage training strategy [15], might not bring additional benefits within our current experimental framework. The VAE-Sampler-Decoder and Direct Variational models outperformed the VAE-Decoder. We speculate that this may be because VAE-Decoder did not utilize the benefits of VI in the reverse training path. The subtle difference between the outputs VAE-Sampler-Decoder and Direct Variational models could be due to the fact that the VAE-Sampler-Decoder attempted to match the z obtained from the $Q(z|y)$ mapper with the underlying z (obtained from the $Q(z|X)$ encoder) forming the saved $P(X|z)$ decoder, whereas the proposed Direct Variational method directly optimized $P(X|z)$ from samples obtained from $Q(z|y)$. Our finding is thus consistent with [16], where the CNN model learned to map the sensor domain data to the image domain information using a single-stage training strategy.

In Figs. 5 & 6, we observed that the E-field reconstruction accuracy is affected by the amplitude profile of the MEPs used as the input to the inverse mapper. The E-field reconstruction was notably better for MEP vectors with larger mean amplitude, variance, and number of muscles with non-zero amplitude. Interestingly, reconstruction error was worse overall but also more variable for stimulations in which only one or two muscles were activated (had non-zero MEP amplitude). As is shown in the "worst" reconstruction (right column) example in Fig 5, single (or few) muscle stimulation can result when the coil is distant from the canonical hand.
area (‘hand knob’) of the motor cortex or when the E-field was relatively low in amplitude and distributed. It is therefore unsurprising that it was challenging for the network to estimate the specific E-field distribution, and it ended up returning a low-intensity distributed profile for these types of stimulations. Such low amplitude MEP response profiles are proximal to the zero activation MEP profile in the vector space. Thus, the set of possible E-field distributions that can produce such MEP responses are in the vicinity of the null space of the transformation matrix equivalent of the cortico-motor mapping for the hand knob area. Future work may build upon the models proposed here to improve E-field estimates for stimuli that induce such small and sparse muscle activation.

The advancement of biophysical modeling of the induced E-field generated by TMS has precipitated efforts to move past ascribing muscle activations induced by TMS to a single point on the scalp. Recently, there have been several concerted efforts to link resulting muscle activation to the complex spatial distribution of the induced electric field [9], [34], [35]. Our model can aid such efforts by providing a framework for describing and testing the assignment of stimulated cortical territory to physiological effects on muscle activity. The proposed model could also prove useful to efforts to optimize coil position and mapping efficiency [36], by generating the optimal E-field distribution for activating a muscle or muscles of interest. In fact, the generative quality of the model, the ability to generate an E-field distribution for a novel muscle activation vector, may enable new investigations into the organization of muscle grouping (sometimes referred to as synergies or modules) on the cortex [10], [37].

There are several limitations of this work. For one, the E-fields we attempted to reconstruct from the MEPs were themselves simulated and calculated using numerical procedures from the coil position and orientation parameters. Thus it would be useful to add to the current procedure an additional step in the inverse calculation that tries to reconstruct the coil parameters. This is a subject of our current work. Secondly, the data set constituted only three subjects. Results from more subjects are clearly needed to validate the robustness of the proposed model. Finally, although we tested five different CNN architectures, there may be yet another architecture that would perform even better.

V. CONCLUSION AND FUTURE WORK

In this work, five 3D CNN models were systematically designed to estimate TMS-induced E-field distributions on the BA4 motor cortex from resultant muscle activation measured as MEPs in an inverse imaging task. Our Direct Variational model, which directly optimized the latent space from both the MEP input and the E-field output during training, emerged as the best performing model, and thus our candidate of choice for M2M-InvNet. In particular the Direct Variational performed better than our other four approaches on all three metrics of evaluation; it showed the lowest root mean square error, the highest average fidelity reconstruction, and the smallest average shift in the center of gravity of the induced fields, when compared to the ground truth. Subsequent examination of the M2M-InvNet inference at different levels of stimulation intensity revealed that both the location and intensity of the stimulation in the target area had substantial impacts on the reconstruction performance, and the number of muscles activated and the mean and variance of their MEPs all generally correlated positively (up to a threshold, with the number of active muscles) with performance.
In future work, we plan to include cortical motor topography mapping using active learning [23] and to study the generation of the volume conductor model by deep learning, to determine if we can combine these with the current expert user-guided mapping and the segmentation-finite element simulation pipeline, respectively, or perhaps even replace either or both entirely.

References


---

In future work, we plan to include cortical motor topography mapping using active learning [23] and to study the generation of the volume conductor model by deep learning, to determine if we can combine these with the current expert user-guided mapping and the segmentation-finite element simulation pipeline, respectively, or perhaps even replace either or both entirely.