# Model Comparison Metrics Require Adaptive Correction If Parameters Are Discretized: Proof-of-Concept Applied to Transient Signals in Dynamic PET

Heather Liu and Evan D. Morris

***Abstract*— Linear parametric neurotransmitter PET (lp-ntPET) is a novel kinetic model that estimates the temporal characteristics of a transient neurotransmitter component in PET data. To preserve computational simplicity in estimation, the parameters of the nonlinear term that describe this transient signal are discretized, and only a limited set of values for each parameter are allowed. Thus, linear estimation can be performed. Linear estimation is implemented using predefined basis functions that incorporate the discretized parameters. The implementation of the model using discretized parameters poses unique challenges for significance testing. Significance testing employs model comparison metrics to determine the significance of the improvement of the fit accomplished by including a basis function, i.e. it determines the presence of a transient signal in the PET data. A false positive occurs when the bases overfit data that do not contain a transient component. The number of parameters in a model, *p*, is necessary to determine the degrees of freedom in the model. In turn, p is crucial for the calculation of model selection metrics and controlling the false positive rate (FPR). In this work, we first explore the effect of parameter discretization on FPR by fitting simulated null data with varying numbers of bases. We demonstrate the dependence of FPR on number of bases. Then, we propose a correction to the number of parameters in the model, $p^{eff}$, which adapts to the number of bases used. Implementing model selection with $p^{eff}$ maintains a stable FPR independent of number of bases.**

Manuscript received January 13, 2020; accepted January 20, 2020. Date of publication February 5, 2020; date of current version June 30, 2020. This work was supported in part by the National Institutes of Health under Grant R01DA038709 (Morris). *(Corresponding author: Heather Liu.)*

Heather Liu is with the Department of Biomedical Engineering, Yale University, New Haven, CT 06520 USA, and also with the Department of Radiology and Biomedical Imaging, Yale School of Medicine, Yale University, New Haven, CT 06520 USA (e-mail: heather.liu@yale.edu).

Evan D. Morris is with the Department of Biomedical Engineering, Yale University, New Haven, CT 06520 USA, also with the Department of Radiology and Biomedical Imaging, Yale School of Medicine, Yale University, New Haven, CT 06520 USA, and also with the Department of Psychiatry, Yale School of Medicine, Yale University, New Haven, CT 06520 USA (e-mail: evan.morris@yale.edu).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2020.2969425

***Index Terms*— lp-ntPET, parametric imaging, model comparison, goodness of fit, basis functions, constrained optimization.**

## I. Introduction

POSITRON emission tomography (PET) makes it possible to image molecular targets with high specificity. Kinetic models are necessary to quantify physiological properties of the target and to describe tracer-target dynamics. The linear-parametric neurotransmitter model (lp-ntPET) estimates the timing of transient neurotransmitter (NT) release occurring within a single scan [1]–[7]. The model has been applied successfully to characterize the dynamics of smoking-induced dopamine (DA) release [8], [9] as well as mu-opioid receptor occupancy after naloxone administration [10]. Various efforts have been made to refine the model's utility, such as development of nonparametric algebraic methods [11] and incorporation into direct reconstruction of PET data [12].

lp-ntPET is formulated as the sum of two components: 1) the tracer component, which quantifies the steady-state properties of the system, and 2) the NT component, which characterizes a transient NT signal that competes with the tracer. lp-ntPET is the linearized version of the ntPET model [5]. Because of its linear form, lp-ntPET can be used to estimate parameters that describe the tracer and NT components on the voxel level with high computational efficiency—thousands of voxels can be fitted, in just minutes. To implement linear estimation, parameters in the NT component that describe the timing of the transient signal are discretized. A limited set of plausible timing parameter values for these discretized timing parameters are defined before estimation. Each discrete combination of possible timing parameters forms one basis function. All combinations together form a library of 'bases' that represents all candidate timing profiles of the transient signal. Linear fitting for the rest of the parameters in the model is performed for each of the bases. The combination of linear parameters and basis function that produces the best fit is retained as the set of optimal parameters.

lp-ntPET (the "full model") is susceptible to overfitting and false positive detection of a transient NT signal.

Thus, significance testing is essential to control the false positive rate (FPR). Without the NT component, lp-ntPET is identical to the MRTM model (hence, the "restricted model") [13]. Model comparison metrics can be used to evaluate the significance of improvement in the fit by the full model over the restricted model. In essence, these metrics adjudicate the need for including the basis functions during fitting, which in turn indicates the presence of a true positive transient in the PET signal.

Previous work by our group has characterized the performance of different model comparison metrics for comparing the full model to the restricted model [4], [14]. However, the effect of parameter discretization on model comparison and FPR has not yet been explored. Model comparison metrics (and statistical tests of models, in general) expect precise knowledge of the number of parameters in the full model, $p_{full}$. However, a limited (discretized) range of values for the timing parameters cannot fully span their respective parameter spaces. Consequently, those parameters contribute only fractional degrees of freedom to the model. In order to properly implement model selection, we propose that it is necessary to determine the "effective" number of parameters in a model. We expect that the effective number of parameters in the full model, $p_{full}^{eff}$, reflects the fractional degrees of freedom contributed by the timing parameters such that $p_{full}^{eff} < p_{full}$. We hypothesize that $p_{full}^{eff}$ depends on the number of bases provided for fitting; the greater the number of distinct bases, the more of parameter space is covered, and the greater the apparent degrees of freedom in the full model. Implementing model comparison with $p_{full}^{eff}$ should alleviate the dependence of FPR on the size of the basis library.

To address our hypothesis, we use simulations of null PET data to demonstrate the dependence of FPR on the number of bases. We then use the demonstrated relationship between number of bases and FPR to determine $p_{full}^{eff}$, the correction to the number of parameters in the full model for a particular number of bases. This correction adapts the number of parameters in the full model to achieve a uniform FPR during model selection independent of number of bases. We evaluate the ability of $p_{full}^{eff}$ to remove the dependence of FPR on number of bases in dynamic 3D and 4D phantom data. Finally, we assess the performance of $p_{full}^{eff}$ in human data from null [$^{11}$C]Raclopride scans, and demonstrate the maintenance of a stable FPR in real data.

## II. Theory

### A. The lp-ntPET Model

lp-ntPET (1a), the "full model", is a multilinear compartmental model containing a time-varying term that describes a transient NT signal (1b). The model is composed of a tracer component and a NT component. The NT component characterizes the effect of a transient time-varying NT signal with $\gamma$, the peak amplitude, and timing parameters: $t_D$, the signal start time relative to injection, $t_P$, the peak time relative to injection, and $\alpha$, the decay rate. These timing parameters are discretized in our linear implementation. u(t) is the unit step function.

$$C_T(t) = R_1 C_R(t) + k_2 \int_0^t C_R(u)\,du - k_{2a} \int_0^t C_T(u)\,du - \gamma \int_0^t C_T(u)\,h_i(u)du \quad (1a)$$

where,

$$h_i(t) = \left(\frac{t-t_D}{t_P-t_D}\right)^{\alpha} exp\left(\alpha\left(1-\frac{t-t_D}{t_P-t_D}\right)\right) u(t-t_D) \quad (1b)$$

The tracer component is composed of the three time-invariant parameters representing kinetic constants describing the tracer. This component is identical to the MRTM [13] (2), the "restricted model".

$$C_T(t) = R_1 C_R(t) + k_2 \int_0^t C_R(u)\,du - k_{2a} \int_0^t C_T(u)\,du \quad (2)$$

$C_T$ and $C_R$ are the concentrations of tracer in the target and reference compartments, respectively. The target compartment contains the NT signal. The reference compartment is used as a proxy for the input function of tracer introduced into the system. $R_1$ is the ratio of tracer delivery to the target and reference compartments; $k_2$ is the efflux rate related to the free diffusion of tracer ; $k_{2a}$ is the efflux rate that incorporates the effects of specific binding of the tracer to the target.

In the basis function implementation of the full model, the NT timing parameters are restricted to a predetermined set of discrete values. All possible combinations of predetermined $t_D$, $t_P$, and $\alpha$ values form a unique library of bases that may vary in both number and timing characteristics. Each basis function is generated by (1b) and then incorporated into (1a) to produce a fit and resultant sum of squared errors (SSE). The combination of the basis function and linear parameters that produces the lowest SSE is retained as the final set of estimated parameters.

### B. Controlling False Positives With Model Comparison Metrics

Model selection metrics evaluate the significance of the improvement in fit that is achieved by including additional parameters in a model. In this case, the metrics evaluate the advantage given by including the NT component in (1a). When used to fit data without an NT signal (null data), basis functions are, by definition, extraneous and any improvement in fit given by the NT component is therefore overfitting. A false positive is defined in this context as a fit to null data for which the full model is erroneously selected over the restricted model. We examined the behavior of three common model selection metrics in determining false positives: the F-statistic (3), the corrected Akaike Information Criterion (*AICc*) (4) [15], and the Bayesian Information Criterion (*BIC*) (5) [16].

$$F = \frac{(SSE_{res} - SSE_{full})/(p_{full} - p_{res})}{(SSE_{full})/(n - p_{full})} \quad (3)$$

$$AICc = 2p + n * \ln\left(\frac{SSE}{n}\right) + \frac{2p^2 + 2p}{n - p - 1} \tag{4a}$$

We define:

$$\Delta AICc = AICc_{full} - AICc_{res} \tag{4b}$$

$$BIC = p * ln\,(n) + n * ln\left(\frac{SSE}{n}\right) \tag{5a}$$

We define:

$$\Delta BIC = BIC_{full} - BIC_{res} \tag{5b}$$

Subscripts *"full"* and *"res"* indicate the full and restricted models, respectively. *SSE* is the sum of squared errors from the fit; $p$ is the number of parameters in the model; $n$ is the number of data points being fitted, i.e. the number of frames per scan.

The criteria for selecting the full model over the restricted model are as follows: 1) the F-statistic must surpass the $F_{critical}$ threshold at the 5% significance level (for the typical threshold of p = 0.05). The $F_{critical}$ threshold is determined by $p_{full} - p_{res}$ degrees of freedom in the numerator and $n - p_{full}$ degrees of freedom in the denominator. 2) $\Delta AICc$ must be less than zero. 3) $\Delta BIC$ must be less than zero.

### C. False Positive Rate and "Effective" Number of Parameters

The false positive rate (FPR) is the fraction of the total number of null data sets, $k$, for which the full model is determined to be superior by a given metric. FPR is defined for each model comparison metric, respectively, as:

$$FPR_F = \frac{\sum_{i=1}^{k}\left[F_i > F_{crit,0.95}\right]}{k} \tag{6}$$

$$FPR_{AICc} = \frac{\sum_{i=1}^{k}\left[\Delta AICc_i < 0\right]}{k} \tag{7}$$

$$FPR_{BIC} = \frac{\sum_{i=1}^{k}\left[\Delta BIC_i < 0\right]}{k} \tag{8}$$

Traditionally, all parameters implied within the right-hand-sides of (6)-(8) are assumed to be known. According to our hypothesis however, $p_{full}$ should actually be a variable that increases with greater coverage of parameter space, i.e., inclusion of more distinct bases. If we stipulate a constant FPR on the left-hand-sides of (6)-(8) and explicitly solve for $p_{full}$ in each case, we obtain the "effective" number of parameters for a specific implementation of the full model. We will denote this unknown variable as $p_{full}^{eff}$. Put simply, the known and unknown variables are switched between the calculations of FPR and $p_{full}^{eff}$. $p_{full}^{eff}$ can then be determined numerically or analytically. $p_{full}^{eff}$ is essentially a modified $p_{full}$ that recalibrates the distribution of each model comparison metric such that a desired FPR of null instances surpasses the critical threshold ($F_{crit}$, for the F-statistic; 0, for $\Delta AICc$, and $\Delta BIC$). We will refer to $F, \Delta AICc$, and $\Delta BIC$ calculated with $p_{full}^{eff}$ in (3)-(5), as $F^{eff}, \Delta AICc^{eff}$, and $\Delta BIC^{eff}$, respectively. These adapted model selection metrics account for the number of bases used during implementation of the full model.

We expect $p_{full}^{eff}$ to lie between 4 and 7 because, of the 7 parameters in the full model, $\{R_1,\ k_2,\ k_{2a},\gamma\}$ are continuous and $\{t_d,\ t_p,\alpha\}$ are discretized. The 4 continuous parameters span their parameter spaces because they are explicitly calculated. Thus, notwithstanding correlation, they contribute 4 full degrees of freedom to the model. The 3 discretized parameters, contained within the basis functions, cannot span their parameter spaces and therefore contribute only fractional degrees of freedom. Taken together, the 3 discretized parameters contribute up to, but less than, 3 additional degrees of freedom to the model.

## III. Methods

### A. Simulations of Ideal Null Data

Noiseless striatal time-activity curves (TACs) were simulated using the simplified reference tissue model (SRTM) [17] to represent [$^{11}$C]Raclopride uptake in the striatum and cerebellum. For the striatum, SRTM parameters were set to: $R_1 = 1$, $k_2 = 0.42$ min$^{-1}$, $BP_{ND} = 3$. A noiseless cerebellum curve was simulated using the 1-tissue compartment model ($K_1 = 0.0918$ mL/(min g), $k_2 = 0.4484$ min$^{-1}$). The arterial input function was taken from a human scan (from Siemens HRRT) following bolus injection of 20 mCi into a male subject (85.45 kg). The noiseless cerebellum curve was taken as the reference region input, $C_R(t)$. Simulated data were binned into 1-minute frames for the first 10 minutes and 3-minute frames for the remainder of the 90-minute scan. The noiseless data simulated from SRTM were then fitted with MRTM (2). This fitted MRTM curve was taken as the ground truth. Noisy data were then generated by adding homoscedastic Gaussian noise to the noiseless MRTM curve. Ten-thousand TACs were generated at each of 11 noise levels, ranging from region-level to voxel-level noise. These data were considered ideal because the fit by the restricted model to the noiseless data contained zero error. All simulations were implemented in MATLAB software (R2017a, The MathWorks, Inc., Natick, MA) using COMKAT modeling routines [18].

### B. Simulations of Realistic Null Data in 3D and 4D Phantom

Realistic data were simulated using the ntPET model [5] to resemble [$^{11}$C]Raclopride uptake by both the striatum and cerebellum. Kinetic parameters were adapted from Pappata *et al.* [19], Morris *et al.* [20], and Fisher *et al.* [21]. Striatal parameters were set to: $K_1 = 0.07344$ mL/(min g), $k_2 = 0.35872$ min$^{-1}$, $k_{on} = 0.0173$ mL/(pmol min), $k_{off} = 0.1363$ min$^{-1}$, $B_{max} = 100$ pmol/mL, $F_v = 0.04$ mL/mL, $k_{on}^{DA} = 0.25$ mL/(pmol min) and $k_{off}^{DA} = 25$ min$^{-1}$($k_D$ = 100 nM). Basal DA concentration was set to 100 nM, so that 50% of receptors would be occupied at baseline. Cerebellum parameters were set to: $K_1 = 0.0918$ mL/(min g), $k_2 = 0.4484$ min$^{-1}$, and $k_{on}^{DA} = k_{off}^{DA} = 0$. The same arterial input function described for the ideal data was used to simulate realistic data. Ten-thousand striatal TACs with voxel-level noise and a noiseless reference region curve were simulated. Time bins were identical to those of the ideal simulations. Noise was added to the striatal TACs. Noise adhered to a

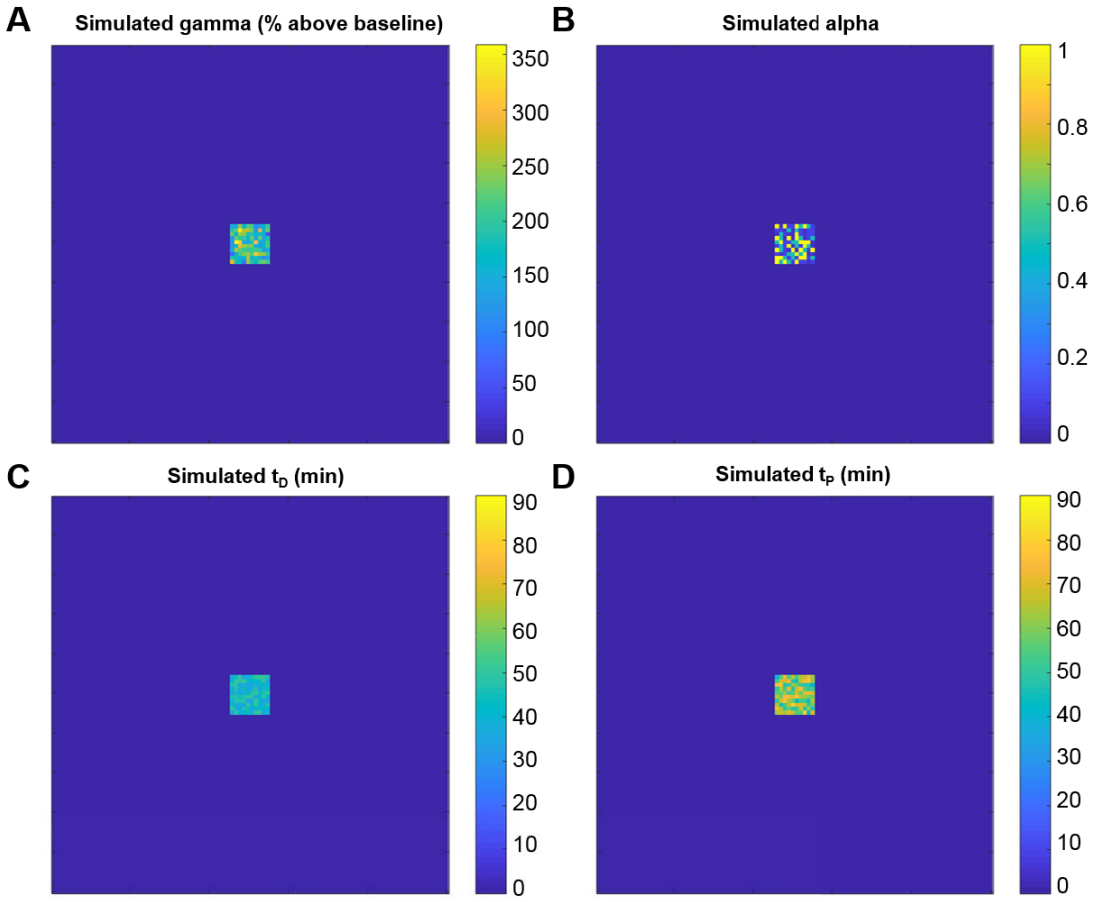


Fig. 1. 2D parametric images of 3D phantom. A) simulated gamma (DA release) values; $\gamma \sim N(200nM,\ 50\ nM)$. B) simulated alpha values; $\alpha \sim U(0.05, 1)$. C) simulated $t_D$ values; $t_D \sim U(35\ min, 50\ min)$. D) simulated $t_P$ values; $t_P \sim U(3\ min, 30\ min)$.

Gaussian distribution with a zero mean and standard deviation modeled according to:

$$\varepsilon_i = noise\ scale * \sqrt{\frac{PET_i \times e^{-\lambda t_i}}{\Delta t_i}} \times e^{\lambda t_i} \tag{9}$$

where $PET_i$ is the signal at a single time point, $i$, without decay correction; $\lambda$ is the decay constant for $^{11}C$; $\Delta t_i$ is the duration of the time frame; $\varepsilon_i$ is the standard deviation of the additive error in the TAC, which was scaled to voxel level noise [22].

Of the 10,000 simulated striatal TACs, 9900 were created to be null and 100 were created to be positive. The positive TACs contained randomly generated time-varying components that adhered to (1b). The timing parameters for the positive simulations were chosen from the following probability density functions: $\gamma \sim$ N(200 nM, 50 nM),$t_D \sim$ U (35min, 50 min) ,$t_P \sim$ U (3 min, 30 min) , $\alpha \sim$ U(0.05, 1); U(min, max) specifies a uniform distribution and N(mean, standard deviation) specifies a normal distribution. Distributions for $t_D$ and $t_P$ were discretized in 3 min intervals. The possible $\alpha$ values were 0.05, 0.1, 0.5, or 1. There were 240 total possible combinations of the timing parameters.

For the 3D phantom (Fig. 1), 10,000 TACs were arranged in a 100 pixel × 100 pixel square; the 100 TACs containing a time-varying component were arranged in a 10 pixel × 10 pixel positive region in the center. Null TACs were assigned to the rest of the phantom.

For the 4D phantom (Fig. 2), 10,000 TACs were arranged in a 50 voxel x 50 voxel × 4 voxel cuboid; the 100 TACs containing a time-varying component were arranged in a 5 voxel × 5 voxel × 4 voxel positive region placed in the center. Null TACs were assigned to the rest of the phantom. This arrangement of voxels was chosen to evoke the 4-slice precommissural striatum mask used in previous studies [9], [23].

### C. Fitting of Ideal Data

Each TAC was fitted with both the full and restricted models. The full model was implemented with libraries of varying numbers of bases, such that there were 10,000 unique sets of fits for each unique combination of noise level and number of bases. For illustration, Fig. 3 shows the libraries with the fewest and most bases. Fitting was implemented using a noiseless $C_T$ in the integral terms in order to eliminate correlated noise between $C_T$ on the left-hand-side and $\int C_T$ on the right-hand-side of (1a) and (2). Although this modification cannot be applied to real data ($C_T$ will never be noiseless), we sought to adhere to the assumptions of linearity as closely as possible in this idealized scenario.

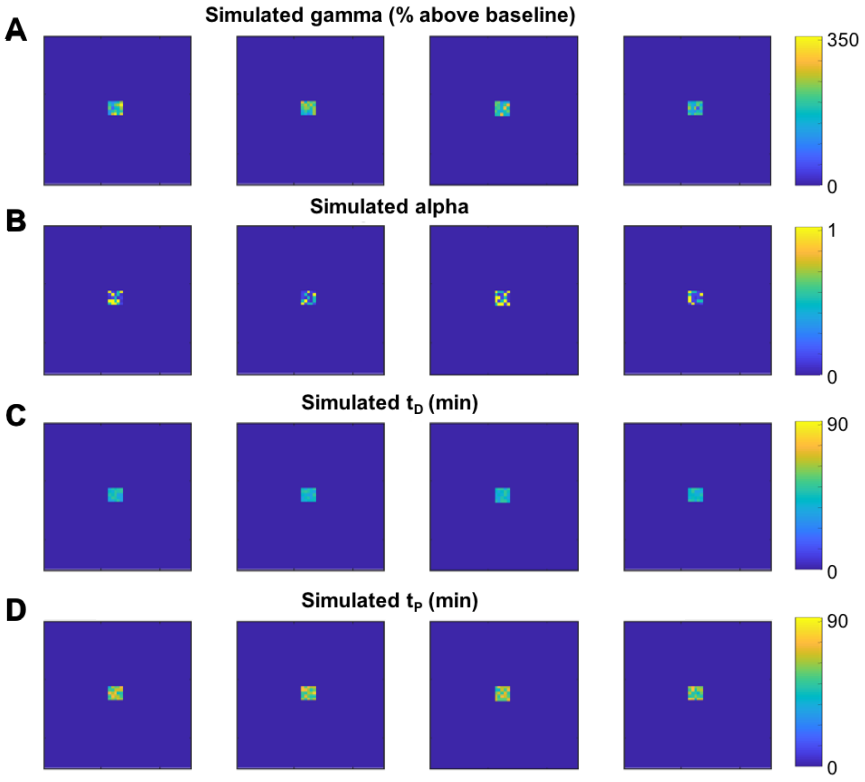


Fig. 2. 3D parametric images of 4D phantom (shown in 4 slices). A) simulated gamma values; $\gamma \sim N$ (200 *nM*, 50 *nM*). B) simulated alpha values; $\alpha \sim U(0.05, 1)$. C) simulated $t_D$ values; $t_D \sim U(35\ min, 50\ min)$. D) simulated $t_P$ values; $t_P \sim U(3\ min, 30\ min)$.

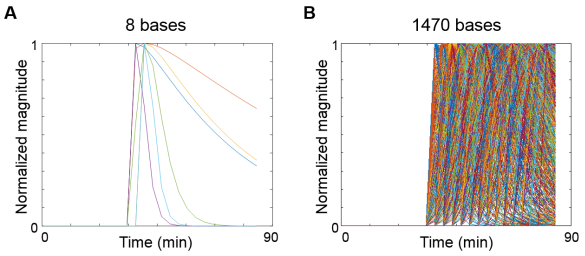


Fig. 3. Response function libraries for fitting with the full model. Fitting libraries varied between A) 6 and B) 1470 bases. Libraries were expanded at ~3 minute resolution, i.e. new $t_D$ and $t_P$ values were appended in 3 minute increments.

## D. Determining FPR and $p^{eff,5\%}$ From Ideal Data

$F, \Delta AICc$, and $\Delta BIC$ were calculated using $p_{full} = 7$ for all fits to the ideal data. First, (6)-(8) were applied to determine FPR as a function of noise and number of bases. Then, FPR was set to 0.05 and (6)-(8) were used to *solve* for $p_{full}^{eff,5\%}$[1] as a variable for every combination of noise and number of bases. $p_{full}^{eff,5\%}$ was determined numerically with the Quasi-Newton algorithm built into MATLAB.

The standard deviations for all FPRs are expected to be small because each FPR is calculated from a large number of data sets (10,000). To confirm, 10 replicates of 10,000 data sets stimulated at voxel noise were fitted with a library of 288 bases (a typical implementation of the model). The standard deviation of the FPR is calculated from the 10 replicates.

[1]For clarity, we have augmented the superscript of $p_{full}^{eff}$ and $\Delta AICc^{eff}$ with, '5%' when referring to values specifically calculated for FPR = 5%

## E. Fitting of Realistic Phantom Data

All phantom TACs were fitted with both the full and restricted models. The full model was implemented with libraries of varying sizes between 9 bases and 240 bases. Libraries varied in resolution of bases, but preserved the range of $\alpha$, $t_D$, and $t_P$ (as shown in Fig. 4) Due to the random nature of the simulated timing parameters, it was necessary to preserve the minimum and maximum limits of each parameter in all fitting libraries. This prevented the estimated value of each timing parameter from being restricted to a range that did not include the parameter's true simulated value. Fitting was implemented with a noisy $C_T$ in the integral terms in (1a) and (2).

$F^{eff,5\%}, \Delta AICc^{eff,5\%}$, and $\Delta BIC^{eff,5\%}$ were calculated for each pair of fits by the full and restricted models, using $p_{full}^{eff}$ (instead of $p_{full}$), as determined for each library size at voxel-level noise. These values for $p_{full}^{eff,5\%}$ are indicated in the blue contour of Fig. 6. Binary "significance masks" were produced to indicate $F > F_{crit}, \Delta AICc < 0$, or $\Delta BIC < 0$ at each pixel or voxel.

$p_{full}^{eff}$ controls for false positives at the voxel level. Due to the violations of linearity and imperfect adherence to model assumptions in real data, FPR tends to be higher by as much as an order of magnitude in real data compared with ideal data. Thus, it is necessary to apply a second level of

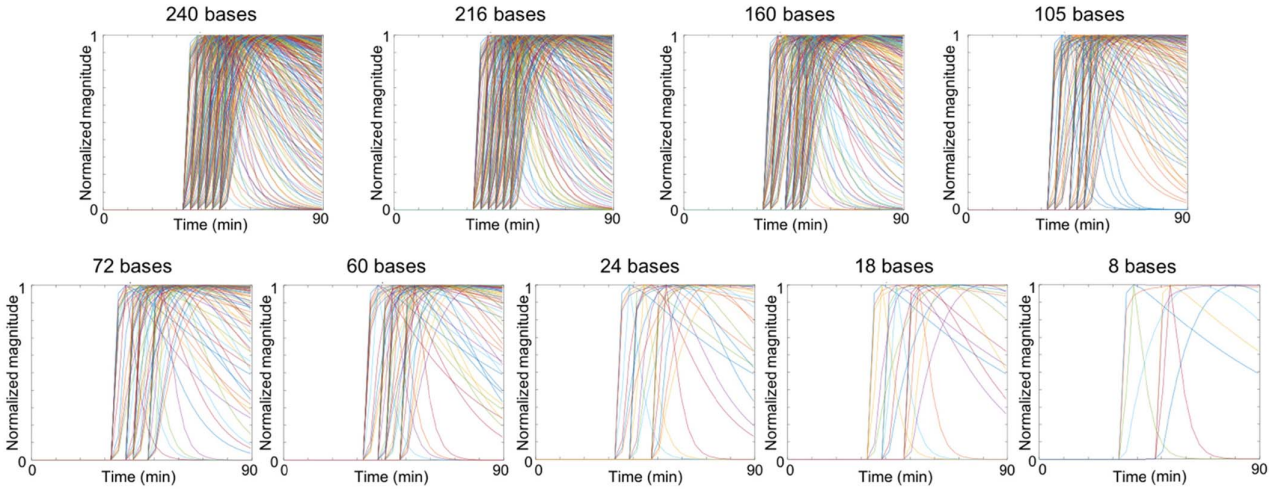


Fig. 4. Response function libraries used for fitting phantom data. As library size increases, the resolution of bases increases but their span is preserved.

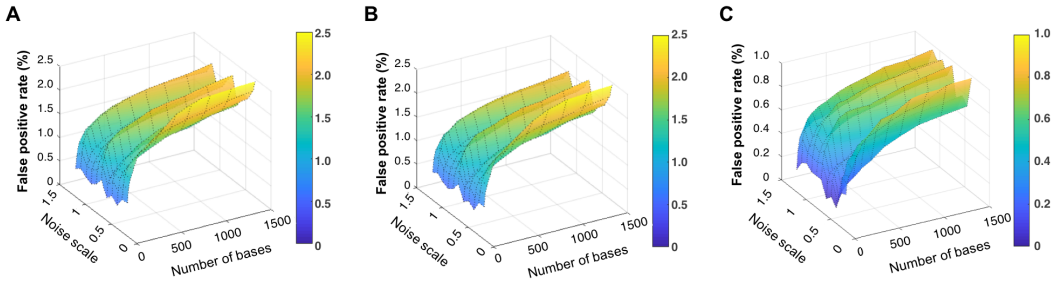


Fig. 5. False positive rate as a function of noise and number of bases for A) F-statistic, B) *ΔAICc* and C) *ΔBIC*. FPR is determined from 10,000 pairs of fits with the full and restricted models, for each unique combination of noise and number of bases. FPR increases with number of bases and appears to saturate. There is no overall trend with noise. All model comparison metrics demonstrate similar overall behavior, but BIC is considerably more conservative.

control to FPR at the image level. Cluster-size thresholding is conventionally used to eliminate false positives at the image level [14]. Various cluster-size thresholds were applied to the significance masks. Cluster-size thresholding is commonly-used as a method to correct for multiple-comparisons in voxel-wise analysis. A single cluster was defined using a blob coloring algorithm based on six-neighborhood connectedness. The FPR was assessed after applying different cluster-size thresholds varying between 1 and 30 pixels/voxels.

### F. Implementation in Human Baseline [$^{11}$C]Raclopride Data

To assess the utility of $F^{eff}$, $\Delta AICc^{eff}$, and $\Delta BIC^{eff}$ in real PET data, all methods described above were applied to dynamic voxel-wise data from two null [$^{11}$C]Raclopride scans. Both subjects were healthy adult male humans. Data from subject 1 (82.1 kg) were acquired following a bolus injection of 13.94 mCi. Data from subject 2 (85.5 kg) were acquired following a bolus injection of 19.73 mCi. No pharmacological or behavioral stimuli occurred before or during either scan. The reference region curve was derived from the cerebellum and was smoothed before fitting. A mask was used to identify 1004 voxels (voxel size: 2 mm × 2 mm × 2 mm, in MNI space) located in the precomissural striatum [23].

## IV. Results

### A. False Positive Rate and "Effective" Number of Parameters in Ideal Simulations

FPR increased at a saturable rate with number of bases for each model comparison metric (Fig. 5). There was no overall dependence of FPR on noise, although some variation of FPR between noise levels can be observed. The standard deviation of the FPR determined for the combination of voxel noise and 288 bases was $1.48 \pm 0.12\%$. All three model comparison metrics demonstrated similar overall behavior. However, $\Delta BIC$ yielded consistently lower FPR than $\Delta AICc$ or $F$. FPR was uniformly below 5% for all model comparison metrics when calculated with $p_{full} = 7$. For space considerations, only results from $\Delta AICc$ will be shown for the remainder of the Results section. The "effective" number of parameters, $p_{full}^{eff,5\%}$, necessary to achieve a FPR of 5% is plotted versus noise level and number of bases in Fig. 6, for $\Delta AICc$. Results for $\Delta BIC$ and $F$ can be found in the Supplemental[2]. $p_{full}^{eff,5\%}$ increased at a saturable rate with number of bases for all metrics. $p_{full}^{eff,5\%}$ varied between 5-6.3, with no dependence on noise.

[2]Supplementary materials are available in the supporting documents /multimedia tab

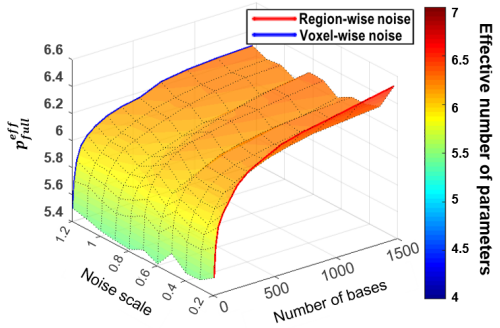


Fig. 6. Surface plot of "effective" number of parameters, $p_{full}^{eff,5\%}$, as determined from $\Delta AICc$. Each combination of number of bases and noise contains the result from analysis of 10,000 pairs of fits. Approximate voxel- and region-level noise are indicated with bolded contours.

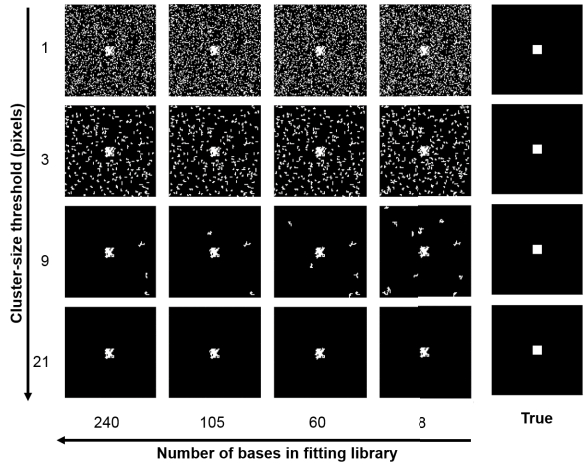


Fig. 7. Select 2D parametric binary images for $\Delta AICc^{eff} < 0$ in 3D phantom for various fitting libraries and cluster-size thresholds. White indicates a pixel for which the full model is determined to be superior to the restricted model. The 10 pixel x 10 pixel positive region is visible in the center of the phantom; the rest of the phantom is null. FPR decreases with increased cluster-size threshold but remains stable across number of bases. At low-resolution libraries (<100 bases), there is a slight inflation of FPR.

### B. Fitting of 3D and 4D Phantom Using $p_{lp-ntPET}^{eff}$

Fig. 7 shows a grid of sample significance masks for the 3D phantom using $\Delta AICc^{eff,5\%}$ at different levels of cluster-size thresholding for voxel-level noise. The background surrounding the center of the phantom became decreasingly noisy as the cluster-size threshold was increased. Fig. 8 shows FPR for the 3D and 4D phantoms, respectively, as a function of number of bases and cluster-size threshold. For both the 3D and 4D phantoms, FPR appears to have been preserved at a constant level across different numbers of bases. FPR decreased with increased cluster-size threshold. A slight inflation of false positives was observed for low-resolution libraries (i.e., fewer bases), compared to all other libraries. This inflation was more prominent in the 4D phantom. In the 3D phantom (Fig. 8A), a cluster-size threshold of 18 pixels eliminated all false positives noncontiguous with the positive region. In the 4D phantom (Fig. 8B), a cluster-size threshold of 30 voxels eliminated nearly all false positives noncontiguous with the positive region.

TABLE I
FPR AS DETERMINED WITH $\Delta AICc^{eff}$ IN HUMAN DATA

| Number of bases in fitting library / Cluster threshold | 30 | 60 | 90 | 204 | 306 |
|---|---|---|---|---|---|
| 1 | 15.4% | 14.4% | 14.7% | 14.8% | 14.7% |
| 3 | 12.1% | 11.3% | 11.9% | 11.8% | 11.9% |
| 9 | 5.7% | 5.4% | 5.6% | 5.0% | 4.6% |
| 15 | 3.1% | 2.9% | 2.9% | 2.8% | 2.3% |
| 24 | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% |

### C. Fitting of Human Baseline [$^{11}$C] Raclopride Data

Application of $\Delta AICc^{eff,5\%}$ for significance testing on human null PET data demonstrated stable FPR across basis libraries of different sizes. Average FPR values (using both subjects) are shown in Table I. Supplemental Table I shows results calculated without the correction ($p_{full} = 7$), for comparison. In this analysis, we defined any voxel for which $\Delta AICc^{eff,5\%} < 0$ as a false positive. Without any cluster-size thresholding, FPR was ~15%. Fig. 9 visualizes the binary significance images for $\Delta AICc^{eff,5\%} < 0$ fitted with 30 bases and thresholded at a cluster-size of 9 voxels, in both subjects. A 9-voxel cluster-size threshold gave an average FPR of ~5% for all fitting libraries. The spatial locations of significant voxels differed between the subjects, suggesting that the significant clusters are, indeed, false positives. All significant voxels were eliminated at a cluster-size threshold of 24 for both subjects. A slightly inflated FPR was produced for libraries <60 bases.

## V. Discussion

We have demonstrated a dependence of FPR on the number of bases used in the implementation of the full lp-ntPET model. To alleviate this dependence, we developed a correction to the number of parameters in the full model, $p_{full}$, yielding a new parameter defining the "effective" number of parameters, $p_{full}^{eff}$. $p_{full}^{eff}$ depends solely on the model and the number of bases. By using $p_{full}^{eff}$ to calculate any standard model comparison metric, the dependence of FPR on number of bases can be eliminated. When applying $p_{full}^{eff,5\%}$ to ideal data, the corrected model comparison metrics yield a consistent FPR of 5%.

### A. FPR and $p_{full}^{eff}$ versus Resolution of Basis Library

Intuitively, the greater the number o bases, the more densely parameter space is covered by the discretized parameters. Without adapting significance testing to properly reflect the greater or lesser coverage of parameter space, the chance of overfitting will be greater or lesser, accordingly. We have demonstrated that this is true by showing that FPR increased with number of bases if 7 parameters were stipulated in the full model. In ideal simulated data, if $p_{full} = 7$, FPR was consistently found to be lower than the expected 5%. In other

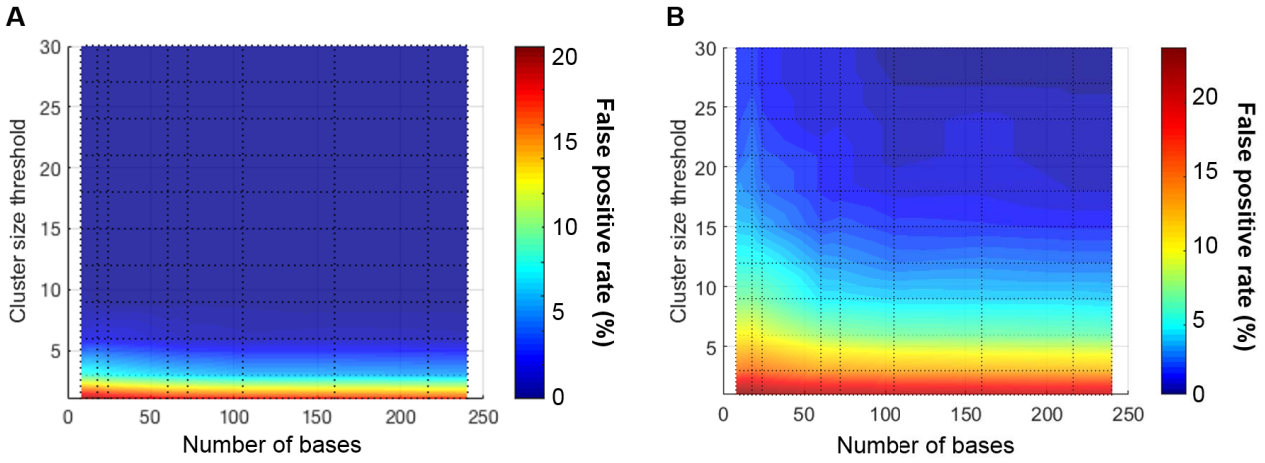


Fig. 8. False positive rate (determined with $\Delta AICc^{eff}$) as a function of number of bases and cluster-size threshold in A) 3D phantom and B) 4D phantom. FPR decreases with increased cluster-size threshold, but decreases more slowly in the 4D phantom, as indicated by the more gradual color gradient moving upwards. FPR remains largely stable across number of bases, with a slight inflation in FPR at low-resolution libraries (<100 bases), apparent in the 4D phantom.

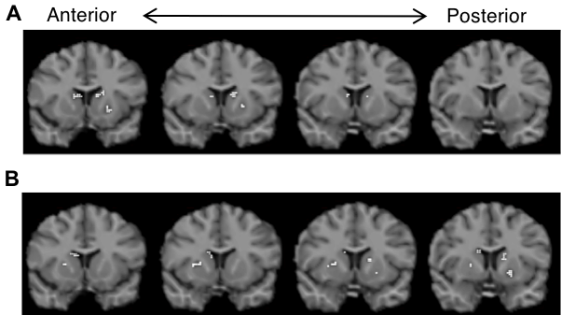


Fig. 9. Binary images for $\Delta AICc^{eff} < 0$ with a cluster-size threshold of 9 voxels in null human PET scans. A) Data for subject 1 fitted with 30 bases. B) Data for subject 2 fitted with 30 bases. White indicates a significant voxel. All 4 contiguous coronal slices of the precommissural striatum are shown in AAL space. Voxel size is 2 mm × 2 mm × 2 mm. FPR is ∼5% for both subjects. The two subjects show different spatial pattern of significant voxels, indicating that activated clusters are false positives.

words, the standard model comparison metrics, $F$, $AICc$, and $BIC$, over-penalized the full model for number of parameters. This indicates that, in fact, the full model behaved as if it contained less than 7 parameters.

Our results confirm our hypothesis that the 3 discretized parameters in the basis function implementation of the full lp-ntPET model assert *fractional* degrees of freedom, and thus $4 < p^{eff}_{full} < 7$. The discretized parameters should not be treated as full parameters during statistical testing because they do not span their full parameter spaces. $p^{eff}_{full}$ adjusts the number of parameters in the full model to reflect the number of bases. As more bases are included, the resolution of the library is increased, and $p^{eff}_{full}$ increases asymptotically towards full apparent degrees of freedom. In practice, model parameters are correlated and are not completely identifiable, so $p^{eff}_{full}$ approaches a value less than 7.

## B. Success of $p^{eff,5\%}_{full}$ in Phantom and Human Data

Realistic data do not perfectly adhere to all assumptions of linearized reference tissue models [24]. The assumption of uncorrelated noise between the dependent variables and the independent variable, $C_T$, is not strictly true. Thus, an FPR that exceeds 5% was expected. Cluster-size thresholding is necessary as a secondary method to control for false positives at the image level. Previous work showed that a cluster-size threshold of 15 voxels was necessary to achieve a 1% cluster-wise FPR (one in 100 activated *clusters* was a false positive) [14]. Here, we chose to define false positives on the voxel level. A cluster-size threshold of ∼9 voxels was sufficient to achieve a 5% voxel-level FPR in a 4D phantom of mostly null voxels. A threshold of 15-18 voxels was necessary to achieve a 1% voxel-level FPR. Comparable cluster-size thresholds did, in fact, achieve similar FPRs in both the human data and the 4D phantom data.

The relationship between decreased FPR and increased cluster-size threshold was nearly identical between the 4D phantom and human data, as seen in the Table I and Fig. 8B. Both phantom and human data demonstrated slightly inflated FPR when the number of bases is <100. This trend is contrary to what was observed without the correction to $p_{full}$ (Fig. 5), and suggests that $p^{eff,5\%}_{full}$ overcompensated for low-resolution libraries. Without cluster-size thresholding, the human data had fewer false positives than the 4D phantom (15% vs. 21%). This can be understood by recognizing that false positives, calculated on the voxel level, were increased by the presence of a true positive cluster in the phantom. This positive cluster acted as a 'seed' for nearby false positive voxels, which allowed them to survive cluster thresholding. In addition, correlation between voxels was not introduced in the simulated data. It is possible that the correlation between voxels in the human data resulted in a lower FPR.

## C. Limitations of the Simulations and the Model

*1) FPR and Noise:* Our initial investigation of bootstrapping the data (data not shown) suggests the ripples seen in Fig. 5 are an artifact of the limited number of simulated data sets. Note that 10,000 simulated curves yields fewer than 500 samples for which $F > F_{crit}$, $\Delta AICc < 0$, or $\Delta BIC < 0$. Thus, these

critical thresholds are essentially determined by less than 5% of samples and are therefore not fully stable.

Theoretically, FPR should not depend on noise level. While it may seem intuitive that increased noise should result in higher FPR, the strength in using model comparison as a method of determining false positives is that it is a "ratio method". As noise increases, the error in the fit increases proportionally for both models. Thus, the ratio of the errors does not change appreciably and the calculated FPR remains fairly stable across noise levels. Note that at zero noise, all model comparison metrics are undefined for null data due to division by zero.

*2) 4D Phantom Data vs. Human Data:* There are some notable differences between the simulated phantom data and human data. In the 4D phantom, the location of the positive cluster was known, and thus easily distinguished from false positives. In the human data, all voxels determined to have a significant time-varying signal were considered to be false positives. Furthermore, noise and time-varying responses in the phantom were generated randomly and independently for each voxel. In real data, some noise correlation is expected as a result of the reconstruction process. Correlation of the timing and amplitude of the biological signal is also expected between neighboring voxels. Both noise and biological correlation would increase the similarity of the TACs in neighboring voxels. As a result, the cluster-size distribution of positive voxels could be skewed.

*3) Other Factors That May Affect Selectivity:* This study explores solely the effect of number of bases on the selectivity of the model. However, other factors, such as time-frame binning, tracer kinetic characteristics, and the selection of the bases themselves, may also affect FPR.

*4) Varying Sensitivity of Discrete Parameters:* The discretized parameters, $t_D$, $t_P$, and $\alpha$, have differential effects on the shape of the basis function. Thus, both the *number* of bases and *which* bases are included in the library will affect the fit to the data by the full model. Not all bases are equally able to describe the noise that occurs in PET data. When varying the size of the libraries, we sought to be as equitable as possible, adding and removing the same number of values for each parameter. However, while $t_D$ and $t_P$ are discretized at the frame resolution, $\alpha$ is continuous and does not affect the timing of the response function in a linear manner. Thus, the incrementing of $\alpha$ values could not be carefully controlled. As a result, it is difficult to eliminate the effect of parameter sensitivity in the incrementing of bases.

Nonetheless, our primary goal was to explore how the size of a discretized parameter space affects the apparent degrees of freedom of the model, i.e. its "effective" number of parameters. The selection of bases is a topic for a separate study that explores the sensitivity of the model to different temporal patterns of true positives.

*5) Selectivity vs. Sensitivity:* This work does not address the sensitivity of the full model. FPR relates only to the *selectivity* of the model. However, there is an indirect relationship. By correcting $p_{full}$, we can increase the resolution of the fitting library without inflating FPR. This could offer an indirect benefit to sensitivity because higher resolution libraries should have better ability to detect true positive signals of varying and unknown temporal profiles.

*6) Broader Implications:* Our observations regarding lp-ntPET and its basis function libraries could have broad implications that extend to other models containing parameters that do not fully span their parameter spaces. SRTM is another example of a kinetic model that is implemented with basis functions for computational efficiency [25]. SRTM is often evaluated against other candidate models for characterization of novel tracers usingmodel comparison metrics [26-29]. Our findings suggest that the number of parameters stipulated for SRTM during model comparison may require a correction based on the number of basis functions used. Spectral analysis, as applied to PET data [30], is also implemented using a limited number of exponential basis functions [31, 32]. While model comparison may not directly apply to spectral analysis, we speculate that the number of basis functions used affects the apparent model degrees of freedom during parameter estimation. Degrees of freedom are not only implicated in model comparison, but in statistical evaluation of linear models, in general.

Beyond our application of interest, we conjecture that any form of constrained optimization may impact a model's apparent degrees of freedom. Non-negative fitting is commonly used to estimate parameters that, by their nature, can only be positive. Other boundaries and constraints placed on an estimated parameter may also limit the parameter's effective degrees of freedom.

*7) Practical Application to Other Data Sets:* The method for determining $p_{full}^{eff}$ presented in this work could be applied to any other tracer and scanner for estimating a transient signal that affects the tracer uptake, which cannot be modeled adequately with time-invariant parameters alone:

1) Use the known tracer kinetic constants, typical measurement variance for the scanner, and an arterial input function for the tracer to generate a large set of "null" data using the appropriate kinetic model and the noise model defined in (9). Null data should be simulated without any signal described by the basis functions.

2) Fit null data with the restricted model and the full model, using basis libraries of varying sizes. When increasing the size of the library, add values for each parameter within a given range, such that for each successive library, the density of sampling increases but the span of the parameter space does not. Add the same number of values to each successive library and space values for each parameter spaced as evenly as possible. Compute the distribution of the desired model comparison metric from the fits to both models.

3) Define the desired FPR on the left-hand-side of (6)-(8). Incorporate $SSE_{full}$, $SSE_{res}$, $p_{res}$, n from each pair of fits into the right-hand-side of (6)-(8). Ideally, the FPR would be selected considering prior information about the model's receiver operating characteristic. The FPR of 5% for this work was selected arbitrarily based on the conventional statistical threshold of $p = 0.05$.

4) Solve for $p_{full}^{eff}$ numerically, for every basis library.

## VI. Conclusion

We have introduced the concept of "effective" number of parameters for models that estimate variables from a discrete set of values. We showed that a discretized parameter contributes only a fractional degree of freedom to the model. The discretized parameter tends towards a full parameter as it is allowed to take on more values. The model comparison process is necessary for controlling false positive results that erroneously indicate the presence of a transient time-varying signal. However, the dependence of FPR on number of bases means that the selectivity of lp-ntPET depends on its implementation; selectivity should depend solely on the noise in the data and the model used. We have developed adaptive model comparison metrics that incorporate $p^{eff}$ to properly account for the coverage of parameter space by the discretized parameters. Applying these adaptive metrics allows for a potential increase in sensitivity without a concomitant decrease in selectivity, as more bases are used.

## Acknowledgment

H. Liu and E. D. Morris would like to thank Dr. Edward Soares for his helpful statistical discussions.

## References

[1] C. C. Constantinescu *et al.*, "Estimation from PET data of transient changes in dopamine concentration induced by alcohol: Support for a non-parametric signal estimation method," *Phys. Med. Biol.*, vol. 53, no. 5, pp. 1353–1367, Mar. 2008.

[2] E. Morris, C. Constantinescu, J. Sullivan, M. Normandin, and L. Christopher, "Noninvasive visualization of human dopamine dynamics from PET images," *NeuroImage*, vol. 51, no. 1, pp. 135–144, May 2010.

[3] M. D. Normandin and E. D. Morris, "Estimating neurotransmitter kinetics with ntPET: A simulation study of temporal precision and effects of biased data," *NeuroImage*, vol. 39, no. 3, pp. 1162–1179, Feb. 2008.

[4] M. D. Normandin, W. K. Schiffer, and E. D. Morris, "A linear model for estimation of neurotransmitter response profiles from dynamic PET data," *NeuroImage*, vol. 59, no. 3, pp. 2689–2699, Feb. 2012.

[5] E. D. Morris *et al.*, "NtPET: A new application of PET imaging for characterizing the kinetics of endogenous neurotransmitter release," *Mol. Imag.*, vol. 4, no. 4, Oct. 2005, Art. no. 7290.2005.05130.

[6] M. D. Normandin and E. D. Morris, "Temporal resolution of ntPET using either arterial or reference region-derived plasma input functions," in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 1, Aug. 2006, pp. 2005–2008.

[7] S. Wang, S. Kim, K. P. Cosgrove, and E. D. Morris, "A framework for designing dynamic lp-ntPET studies to maximize the sensitivity to transient neurotransmitter responses to drugs: Application to dopamine and smoking," *NeuroImage*, vol. 146, pp. 701–714, Feb. 2017.

[8] E. D. Morris *et al.*, "Creating dynamic images of short-lived dopamine fluctuations with lp-ntPET: Dopamine movies of cigarette smoking," *J. Vis. Exp.*, vol. 78, Aug 2013, Art. no. e50358.

[9] K. P. Cosgrove *et al.*, "Sex differences in the brain's dopamine signature of cigarette smoking," *J. Neurosci.*, vol. 34, no. 50, pp. 16851–16855, Dec. 2014.

[10] J. Johansson *et al.*, "Intranasal naloxone rapidly occupies brain mu-opioid receptors in human subjects," *Neuropsychopharmacology*, vol. 44, no. 9, pp. 1667–1673, Aug. 2019.

[11] C. C. Constantinescu, C. Bouman, and E. D. Morris, "Nonparametric extraction of transient changes in neurotransmitter concentration from dynamic PET data," *IEEE Trans. Med. Imag.*, vol. 26, no. 3, pp. 359–373, Mar. 2007.

[12] G. I. Angelis, J. E. Gillam, W. J. Ryder, R. R. Fulton, and S. R. Meikle, "Direct estimation of voxel-wise neurotransmitter response maps from dynamic PET data," *IEEE Trans. Med. Imag.*, vol. 38, no. 6, pp. 1371–1383, Jun. 2019.

[13] M. Ichise *et al.*, "Linearized reference tissue parametric imaging methods: Application to [$^{11}$C]DASB positron emission tomography studies of the serotonin transporter in human brain," *J. Cerebral Blood Flow Metabolism*, vol. 23, no. 9, pp. 1096–1112, Sep. 2003.

[14] S. J. Kim, J. M. Sullivan, S. Wang, K. P. Cosgrove, and E. D. Morris, "Voxelwise lp-ntPET for detecting localized, transient dopamine release of unknown timing: Sensitivity analysis and application to cigarette smoking in the PET scanner," *Hum. Brain Mapp.*, vol. 35, no. 9, pp. 4876–4891, Sep. 2014.

[15] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. AC-19, no. 6, pp. 716–723, Dec. 1974.

[16] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, Mar. 1978.

[17] A. A. Lammertsma and S. P. Hume, "Simplified reference tissue model for PET receptor studies," *NeuroImage*, vol. 4, no. 3, pp. 153–158, Dec. 1996.

[18] R. F. Muzic and S. Cornelius, "COMKAT: Compartment model kinetic analysis tool," *J. Nucl. Med.*, vol. 42, no. 4, pp. 636–645, Apr. 2001.

[19] S. Pappata, "*In vivo* detection of striatal dopamine release during reward: A PET study with [$^{11}$c]raclopride and a single dynamic scan approach," *NeuroImage*, vol. 16, no. 4, pp. 1015–1027, Aug. 2002.

[20] E. D. Morris, R. E. Fisher, N. M. Alpert, S. L. Rauch, and A. J. Fischman, "*In vivo* imaging of neuromodulation using positron emission tomography: Optimal ligand characteristics and task length for detection of activation," *Hum. Brain Mapp.*, vol. 3, no. 1, pp. 35–55, 1995.

[21] R. E. Fisher, E. D. Morris, N. M. Alpert, and A. J. Fischman, "*In vivo* imaging of neuromodulatory synaptic transmission using PET: A review of relevant neurophysiology," *Hum. Brain Mapp.*, vol. 3, no. 1, pp. 24–34, 1995.

[22] B. M. Mazoyer, R. H. Huesman, T. F. Budinger, and B. L. Knittel, "Dynamic PET data analysis," *J. Comput. Assist. Tomogr.*, vol. 10, no. 4, pp. 645–653, Jul. 1986.

[23] D. Martinez *et al.*, "Imaging human mesolimbic dopamine transmission with positron emission tomography. Part II: Amphetamine-induced dopamine release in the functional subdivisions of the striatum," *J. Cereb. Blood Flow Metabolism*, vol. 23, no. 3, pp. 285–300, Mar. 2003.

[24] C. A. Salinas, G. E. Searle, and R. N. Gunn, "The simplified reference tissue model: Model assumption violations and their impact on binding potential," *J. Cerebral Blood Flow Metab*, vol. 35, no. 2, pp. 304–311, Feb. 2015.

[25] R. N. Gunn, A. A. Lammertsma, S. P. Hume, and V. J. Cunningham, "Parametric imaging of ligand-receptor binding in PET using a simplified reference region model," *NeuroImage*, vol. 6, no. 4, pp. 279–287, Nov. 1997.

[26] I. L. Alves *et al.*, "Pharmacokinetic modeling of [($^{11}$)C]flumazenil kinetics in the rat brain," *EJNMMI Res*, vol. 7, no. 1, p. 17, Dec. 2017.

[27] M. Kessler *et al.*, "GABA$_A$ receptors in the mongolian gerbil: A PET study using [$^{18}$F]flumazenil to determine receptor binding in young and old animals," in *Molecular Imaging and Biology*. New York, NY, USA: Springer, May 2019, pp. 1–13.

[28] M. Yaqub *et al.*, "Evaluation of tracer kinetic models for analysis of [$^{18}$F]FDDNP studies," *Mol. Imag. Biol.*, vol. 11, no. 5, pp. 322–333, Sep. 2009.

[29] M. Yaqub *et al.*, "Quantification of dopamine transporter binding using [$^{18}$F]FP-$\beta$-CIT and positron emission tomography," *NeuroImage*, vol. 27, no. 7, pp. 1397–1406, Jul. 2007.

[30] V. J. Cunningham and T. Jones, "Spectral analysis of dynamic PET studies," *J. Cerebral Blood Flow Metabolism*, vol. 13, no. 1, pp. 15–23, Jan. 1993.

[31] J. F. Myers *et al.*, "Characterisation of the contribution of the GABA-benzodiazepine $\alpha$1 receptor subtype to [$^{11}$C]Ro15-4513 PET images," *J. Cerebral Blood Flow Metabolism*, vol. 32, no. 4, pp. 731–744, Apr. 2012.

[32] F. Turkheimer, R. M. Moresco, G. Lucignani, L. Sokoloff, F. Fazio, and K. Schmidt, "The use of spectral analysis to determine regional cerebral glucose utilization with positron emission tomography and [$^{18}$F]fluorodeoxyglucose: Theory, implementation, and optimization procedures," *J. Cerebral Blood Flow Metabolism*, vol. 14, no. 3, pp. 406–422, May 1994.