Approximate Conditional Coverage via Neural Model Approximations

Allen Schmaltz and Danielle Rasooly

Reexpress AI, Inc. and Harvard University

Workshop on Distribution-Free Uncertainty Quantification at ICML (July 2022), Baltimore, Maryland

5 minute overview

re.express

Desiderata for deploying deep learning models

- High accuracy for point predictions
    - Typically necessary but *not sufficient*

Desiderata for deploying deep learning models

- High accuracy for point predictions
  - Typically necessary but *not sufficient*

- Interpretability
  - Map: Relevant feature subsets to those with known labels
- Local updatability
- Uncertainty quantification

We seek a general framework with all of these properties

Desiderata for uncertainty quantification

- Minimize distributional assumptions
- Robustness to data shifts (covariate & label shift)
- Out-of-distribution data does not lead to unexpected catastrophes
- Minimize free parameters (i.e., practical reliability)
- Informative coverage
  - Prefer conservative coverage over under-coverage

Split-conformal prediction sets for classification

- Computationally expensive blackbox: F
- Training dataset: $\mathscr{D}_{\text{tr}} = \{(X_i, Y_i)\}_{i=1}^{I}$ with $Y_i \in \mathscr{Y} = \{1, \ldots, C\}$
- Held-out labeled calibration dataset: $\mathscr{D}_{\text{ca}} = \{(X_j, Y_j)\}_{j=I+1}^{N=I+J}$

Split-conformal prediction sets for classification

- Computationally expensive blackbox: F
- Training dataset: $\mathscr{D}_{\text{tr}} = \{(X_i, Y_i)\}_{i=1}^{I}$ with $Y_i \in \mathscr{Y} = \{1, \ldots, C\}$
- Held-out labeled calibration dataset: $\mathscr{D}_{\text{ca}} = \{(X_j, Y_j)\}_{j=I+1}^{N=I+J}$
- Seek: A prediction set $\hat{\mathscr{C}}(X_{N+1}) \in 2^C$ for a new, unseen test instance X_{N+1} from $\mathscr{D}_{\text{te}}$
  - Contains the true label with coverage level $1 - \alpha \in (0,1)$ *on average*

Split-conformal prediction sets for classification

- Computationally expensive blackbox: F
- Training dataset: $\mathscr{D}_{\text{tr}} = \{(X_i, Y_i)\}_{i=1}^{I}$ with $Y_i \in \mathscr{Y} = \{1, \ldots, C\}$
- Held-out labeled calibration dataset: $\mathscr{D}_{\text{ca}} = \{(X_j, Y_j)\}_{j=I+1}^{N=I+J}$
- Seek: A prediction set $\hat{\mathscr{C}}(X_{N+1}) \in 2^C$ for a new, unseen test instance X_{N+1} from $\mathscr{D}_{\text{te}}$
    - Contains the true label with coverage level $1 - \alpha \in (0,1)$ *on average*
- Finite-sample *marginal* guarantee:
    - $\mathbb{P}\left\{ Y_{N+1} \in \hat{\mathscr{C}}(X_{N+1}) \right\} \geq 1 - \alpha$
    - Via $\hat{\mathscr{C}}(x_{N+1}) = \left\{ c \in \mathscr{Y} : \hat{\pi}^c(x_{N+1}) \geq \hat{\tau}^{\alpha} \right\}$, where $\hat{\tau}^{\alpha} = 1 - \hat{l}^{\alpha}$

Quantile threshold

conditional coverage

- Finite-sample *conditional* coverage:

$$\mathbb{P}\left\{Y_{N+1} \in \hat{\mathscr{C}}(X_{N+1}) \mid X_{N+1} = x\right\} \geq 1 - \alpha$$

conditional coverage

- Finite-sample *conditional* coverage:

$$\mathbb{P}\left\{ Y_{N+1} \in \hat{\mathscr{C}}(X_{N+1}) \mid X_{N+1} = x \right\} \geq 1 - \alpha$$

Approximate conditional coverage

- Finite-sample *conditional* coverage:

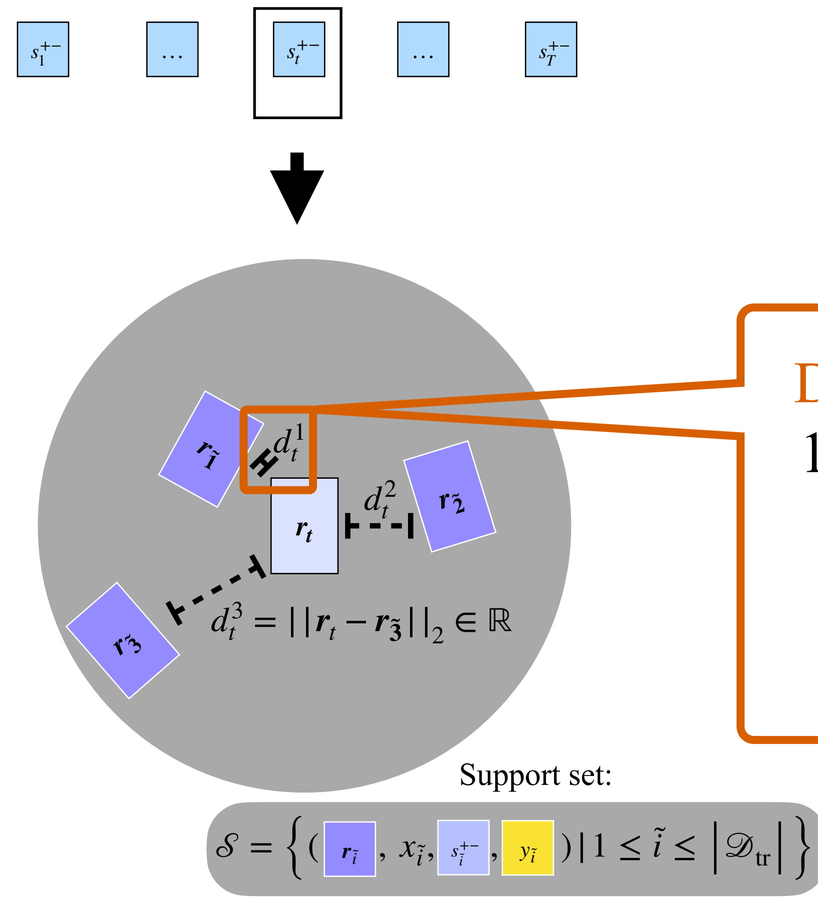
$$\mathbb{P}\left\{ Y_{N+1} \in \hat{\mathscr{C}}(X_{N+1}) \mid X_{N+1} = x \right\} \geq 1 - \alpha$$

- Finite-sample *approximate conditional* coverage:

$$\mathbb{P}\left\{ Y_{N+1} \in \hat{\mathscr{C}}(X_{N+1}) \mid X_{N+1} \in \mathscr{B}(x), Y_{N+1} = y \right\} \geq 1 - \alpha, \text{ with } P_X(\mathscr{B}(x)) \geq \xi$$

Recast a prediction as a weighting over the training set

KNN Approximation

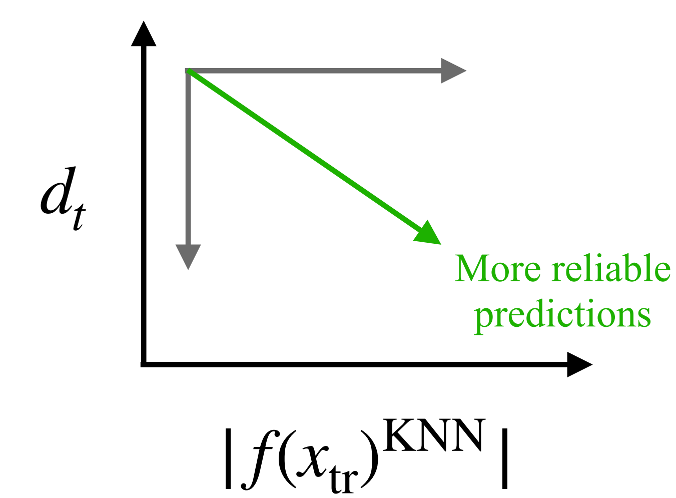


Data uncertainty: Distance to 1st match (d_t), an exogenous factor, captures uncertainty w.r.t. data (training data compared to test data).

Model uncertainty: This bounded value reaches its min/max when $\tanh(s_k^{+-})$ & y (or y_k, with token-level labels) agree, for all k (assuming $\gamma > 0$).

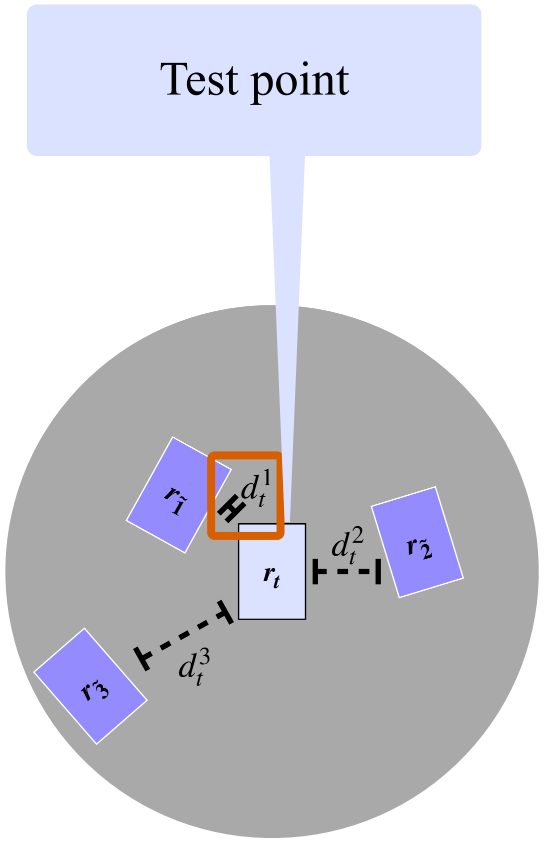
$$
\begin{aligned}
s_t^{+-} \approx \beta &+ w_1 \cdot \left(\tanh(s_{\tilde1}^{+-}) + \gamma \cdot y_{\tilde1}\right) \\
&+ w_2 \cdot \left(\tanh(s_{\tilde2}^{+-}) + \gamma \cdot y_{\tilde2}\right) \\
&+ w_3 \cdot \left(\tanh(s_{\tilde3}^{+-}) + \gamma \cdot y_{\tilde3}\right)
\end{aligned}
$$

$$
w_k = \frac{\exp(-d_k/\eta)}{\sum_{k'=1}^{3} \exp(-d_{k'}/\eta)}
$$

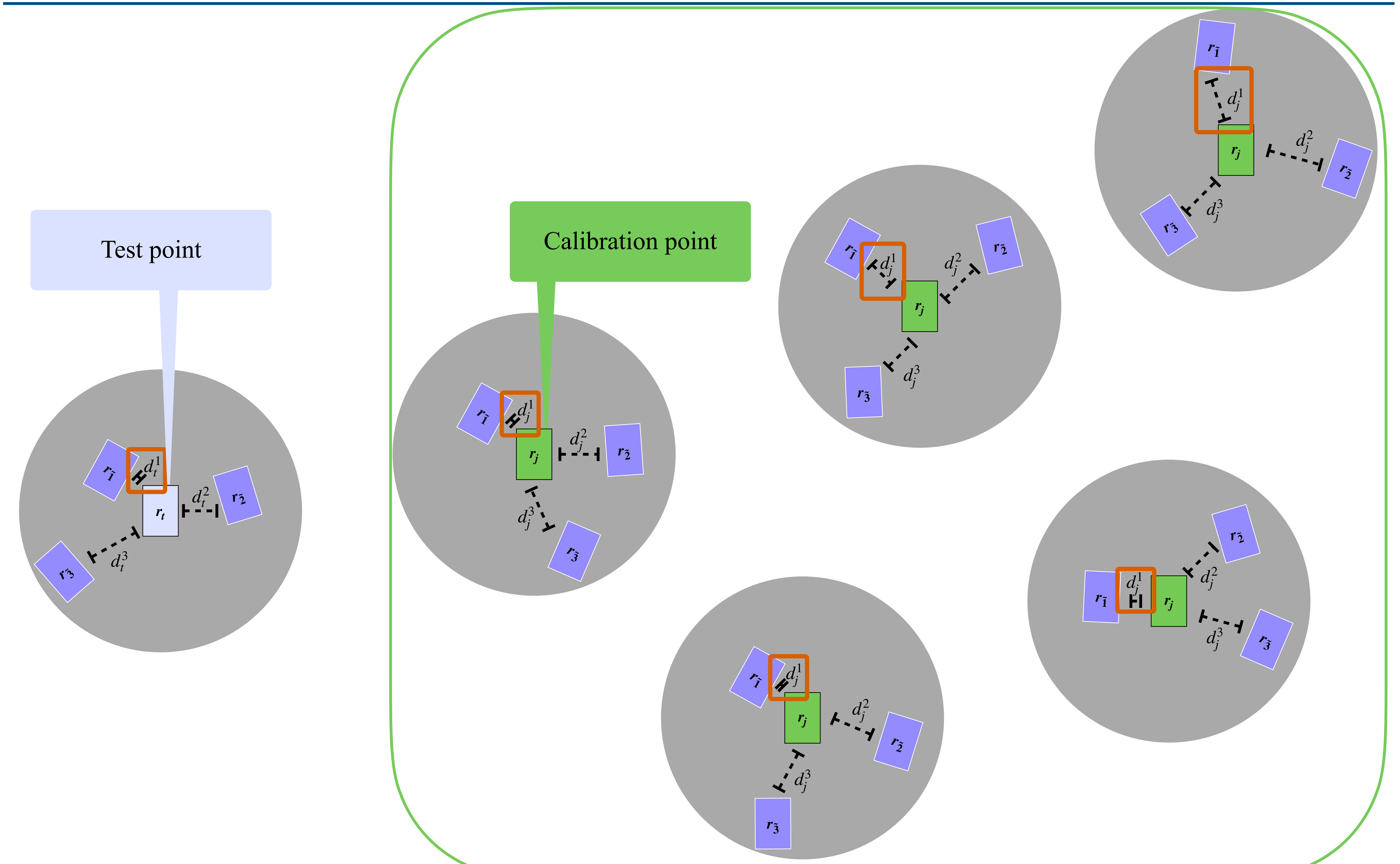


Magnitude of the K-NN Output

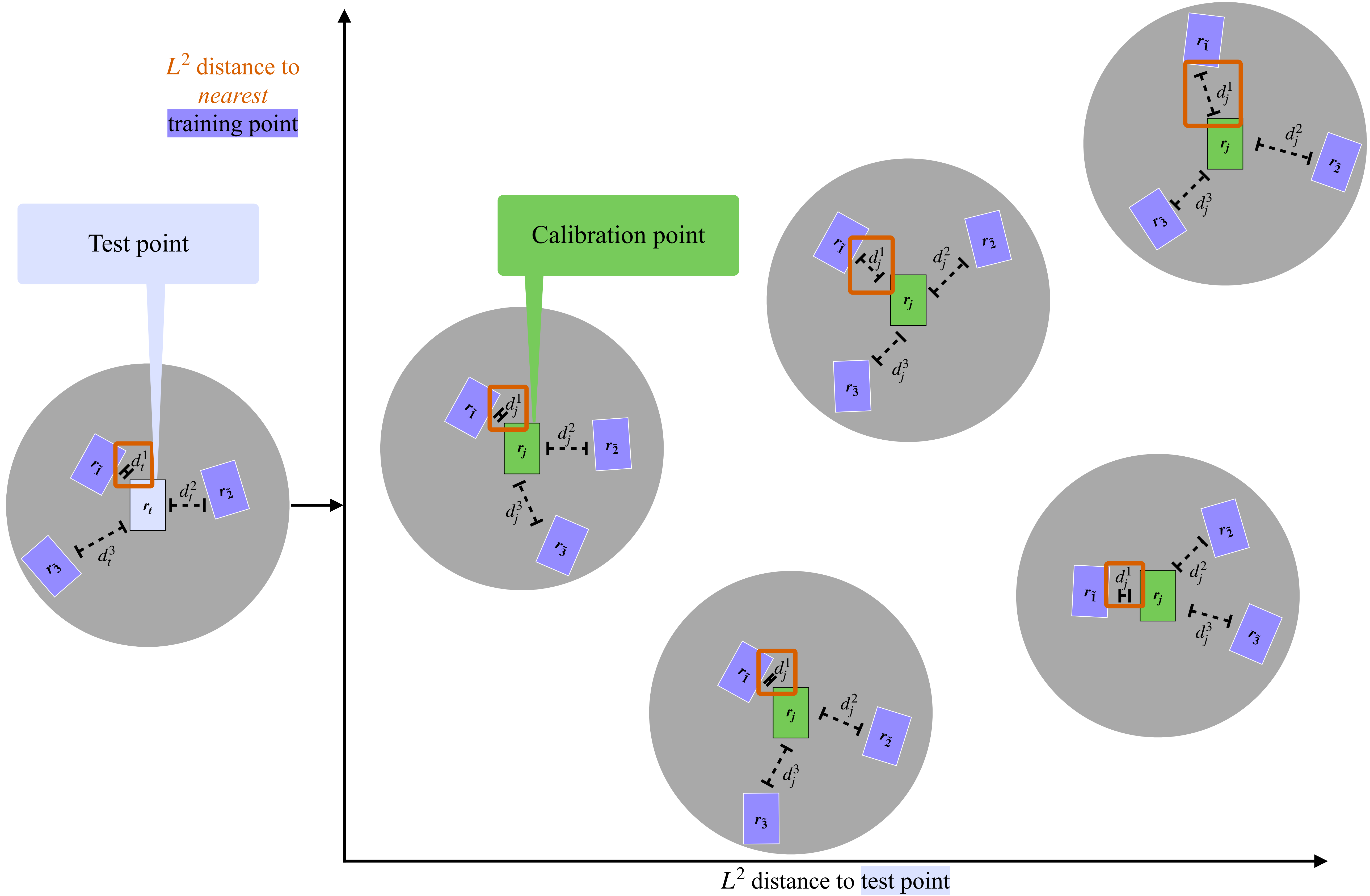
Construct approximation for test point



Construct approximation for test point and all calibration points

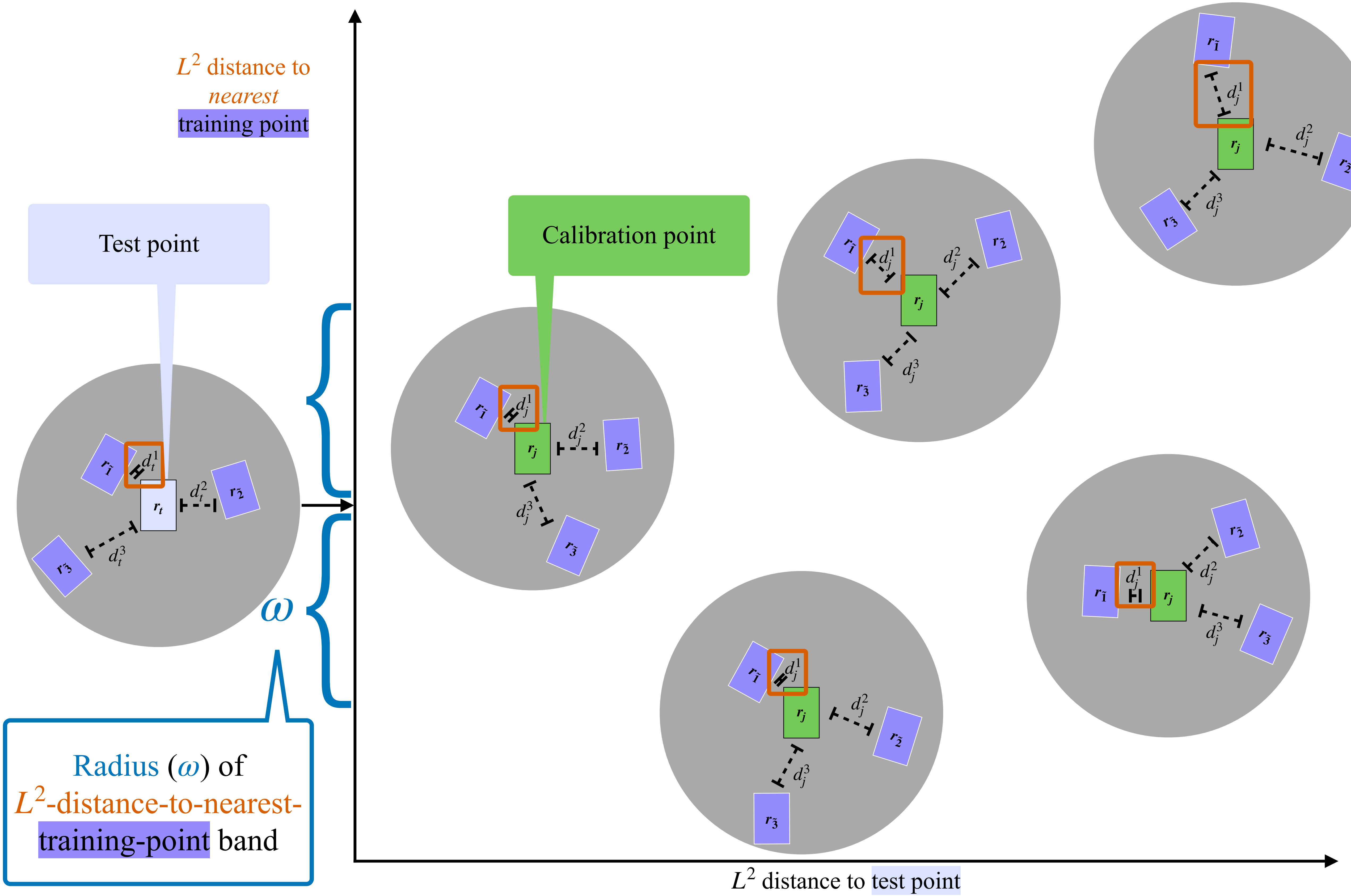


Relate test point to distribution of calibration points

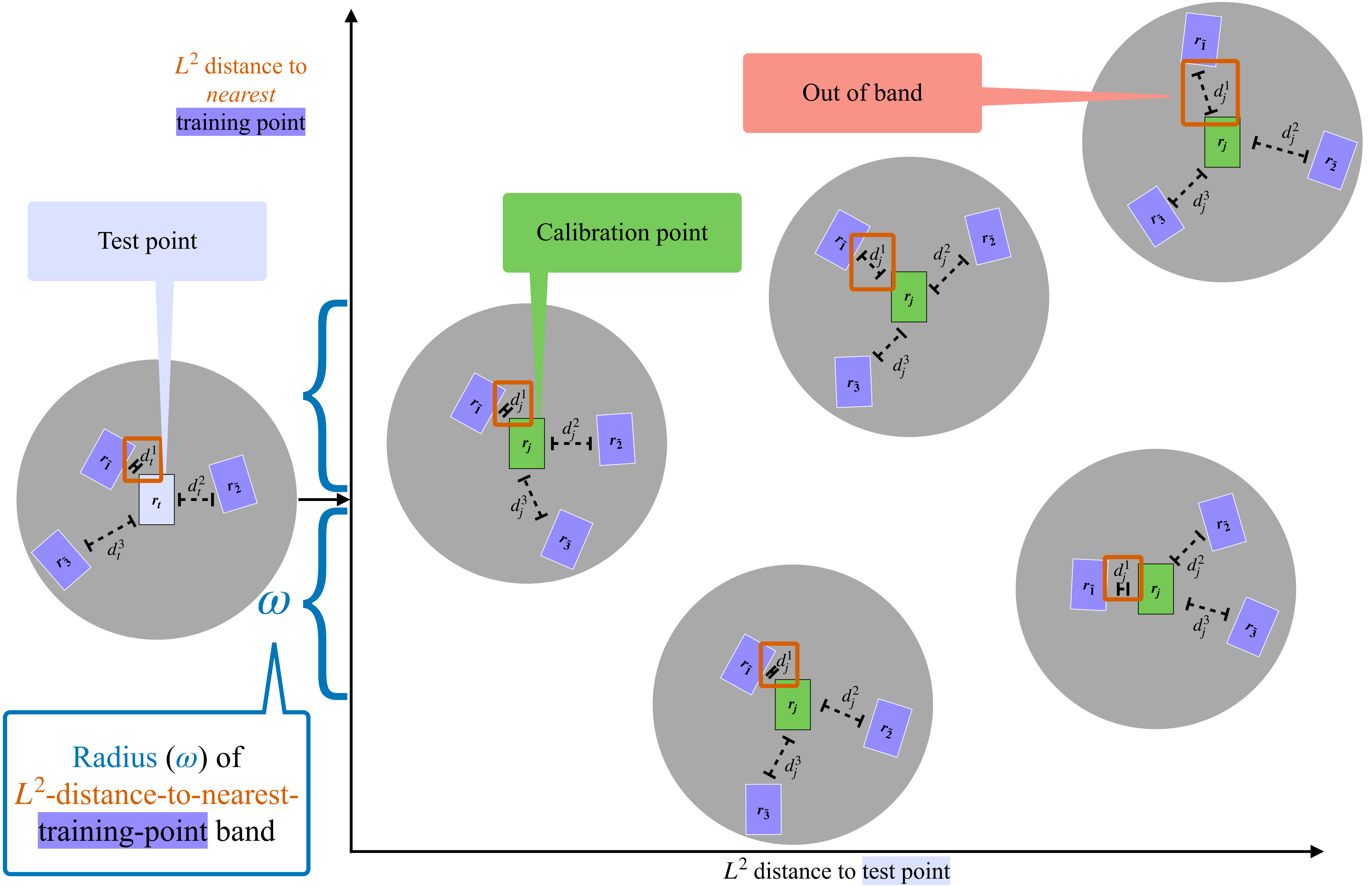


re.express

Construct a distance band around the test point

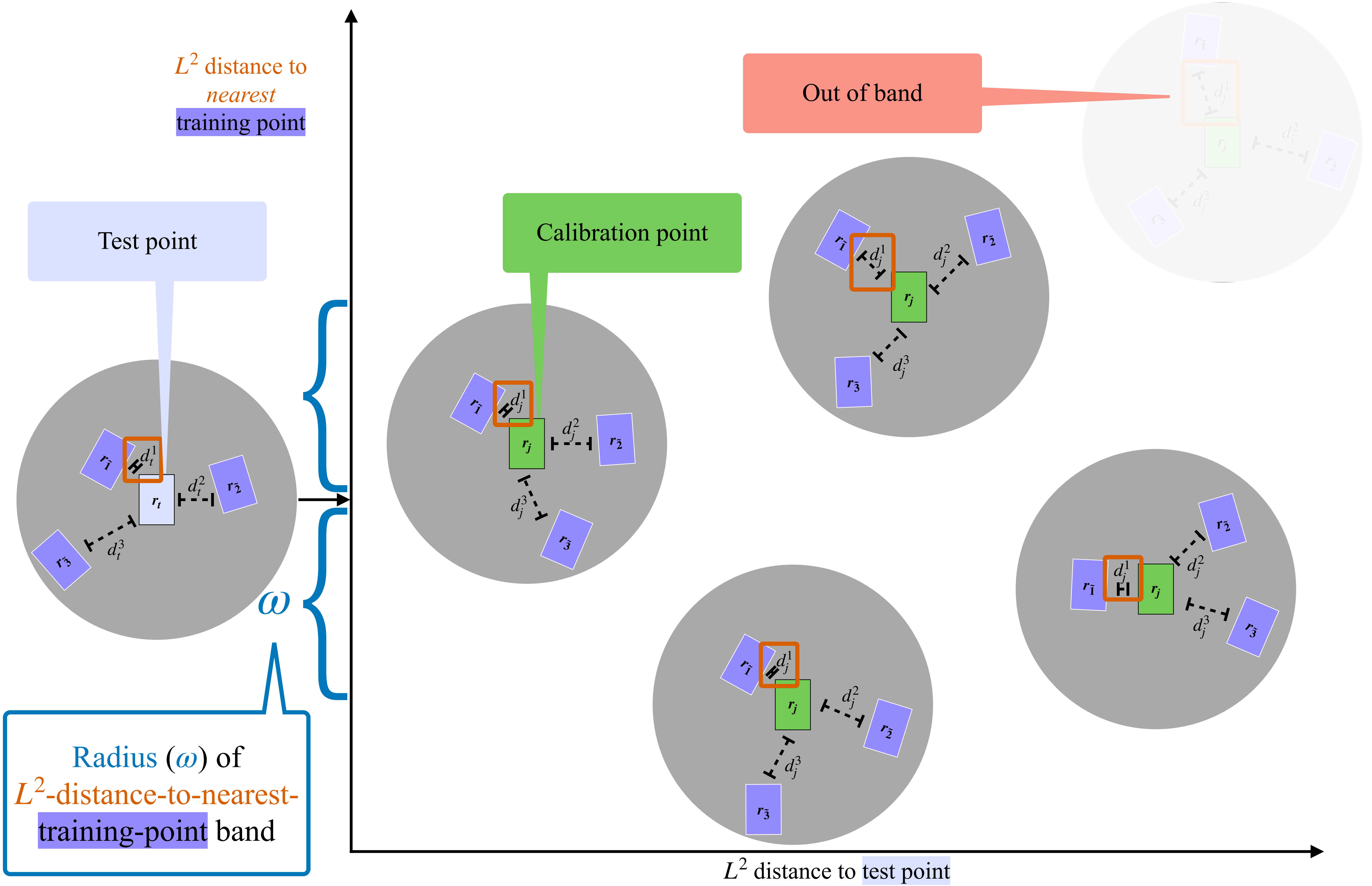


Constrain calibration points to distance band

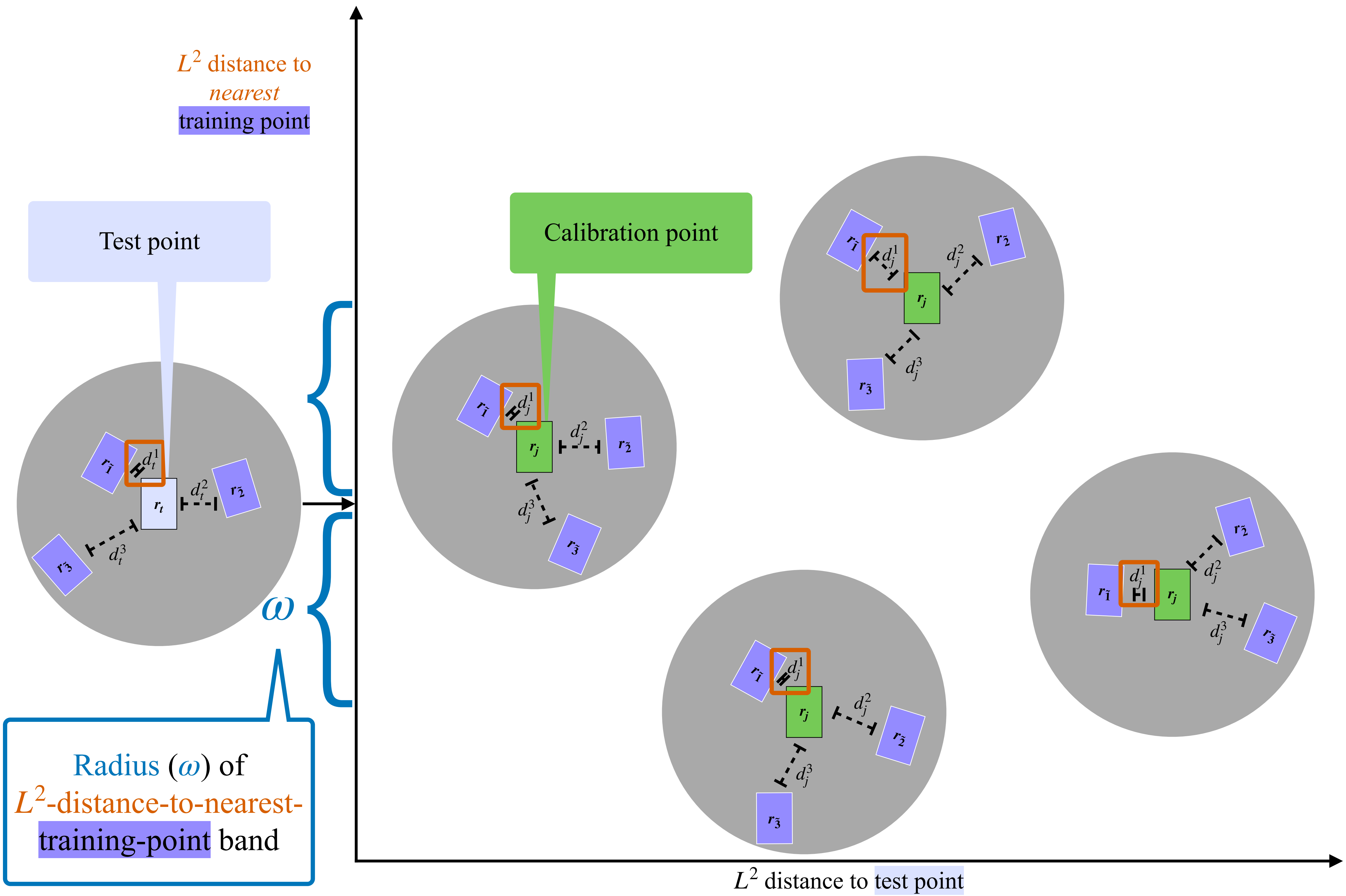


re.express

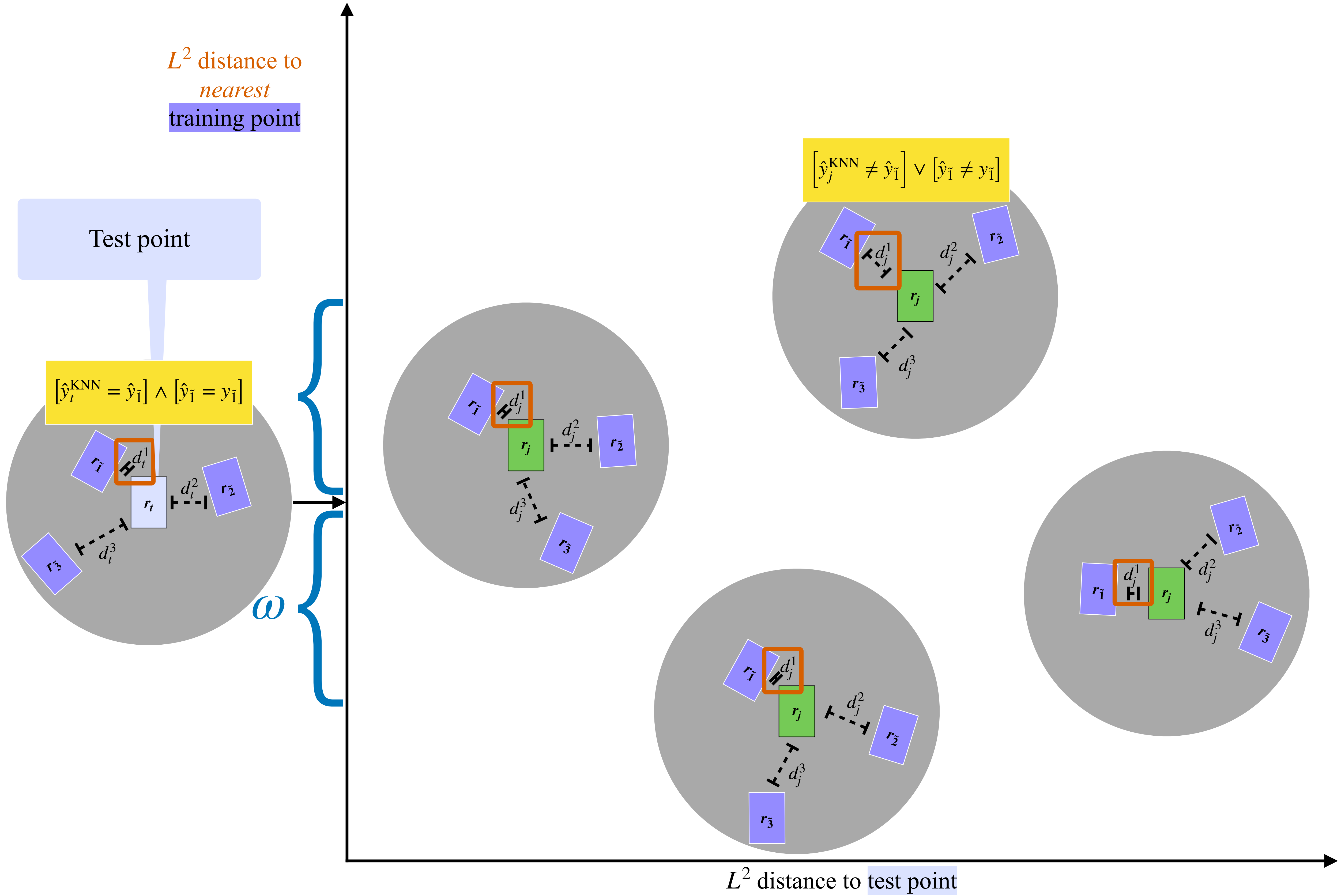
Constrain calibration points to distance band



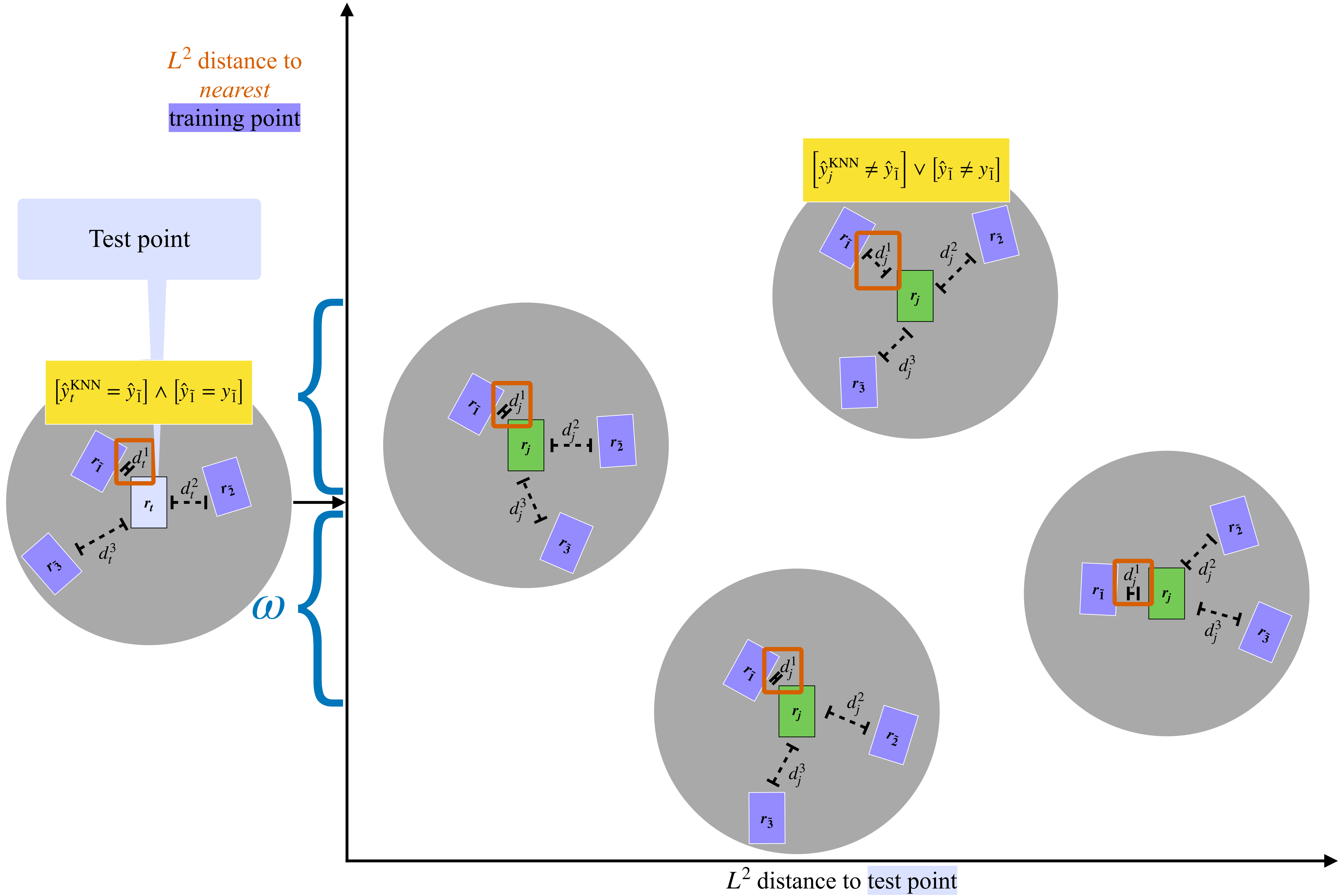
Constrain calibration points to distance band



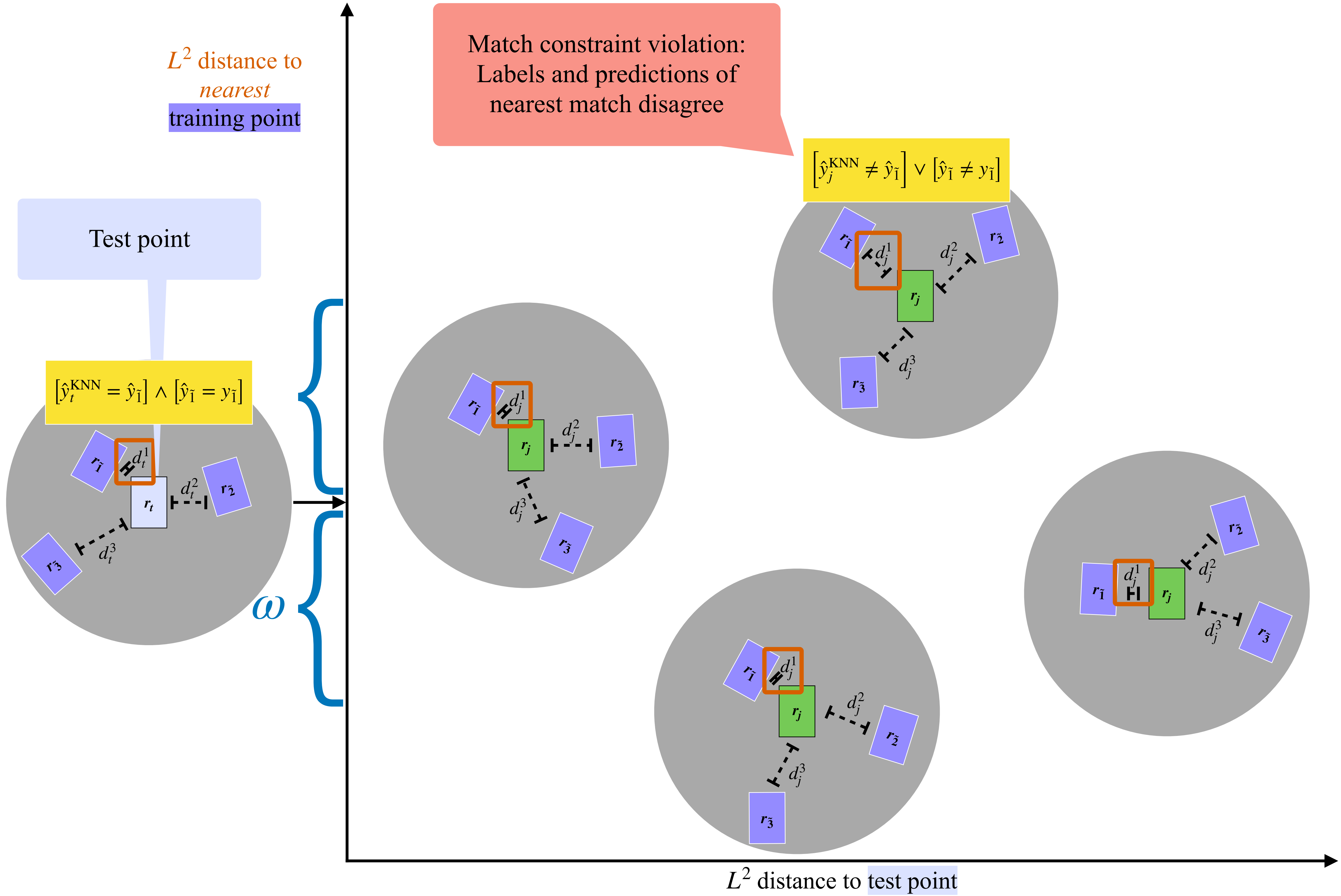
match constraint feature



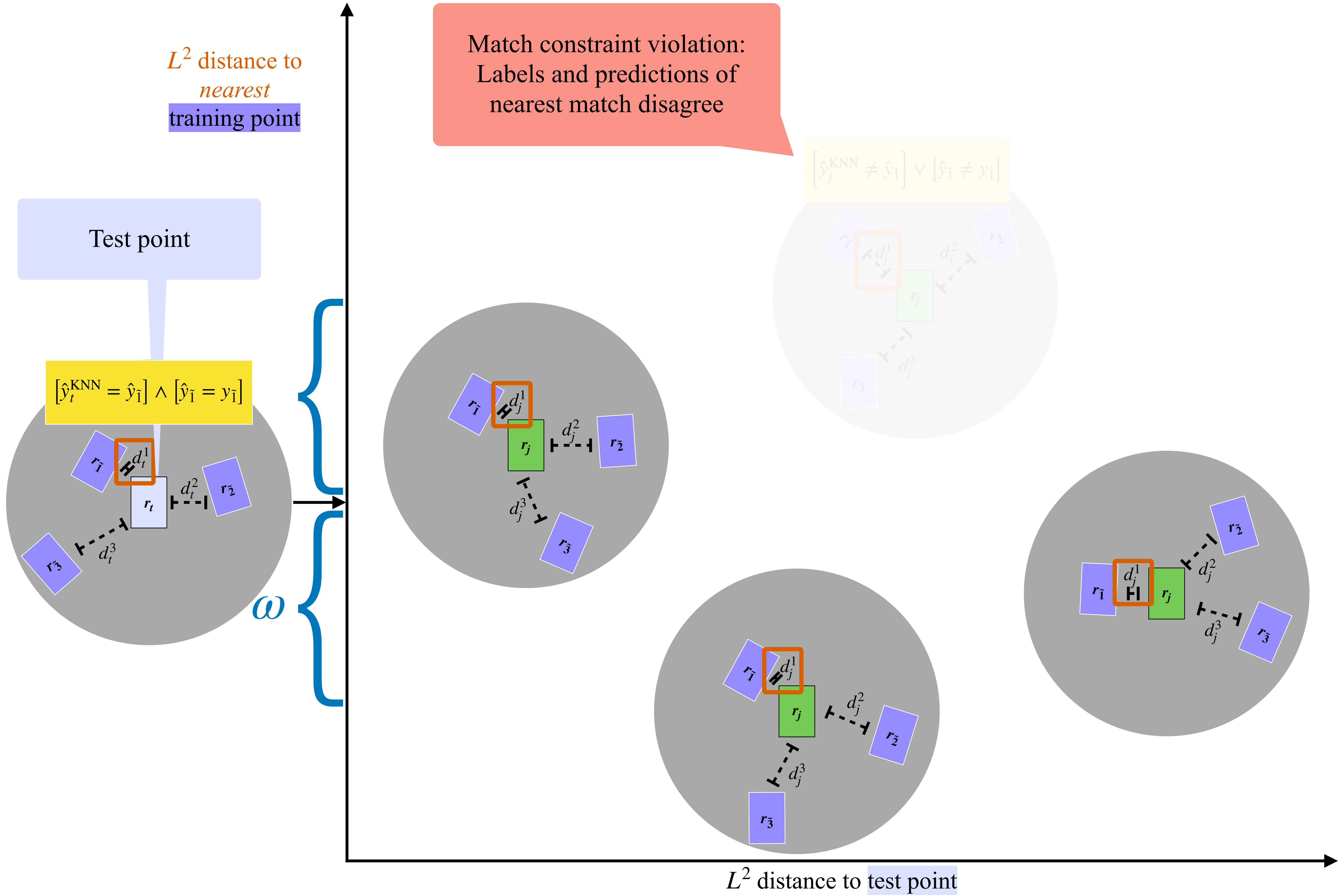
Constrain calibration points to match constraint feature



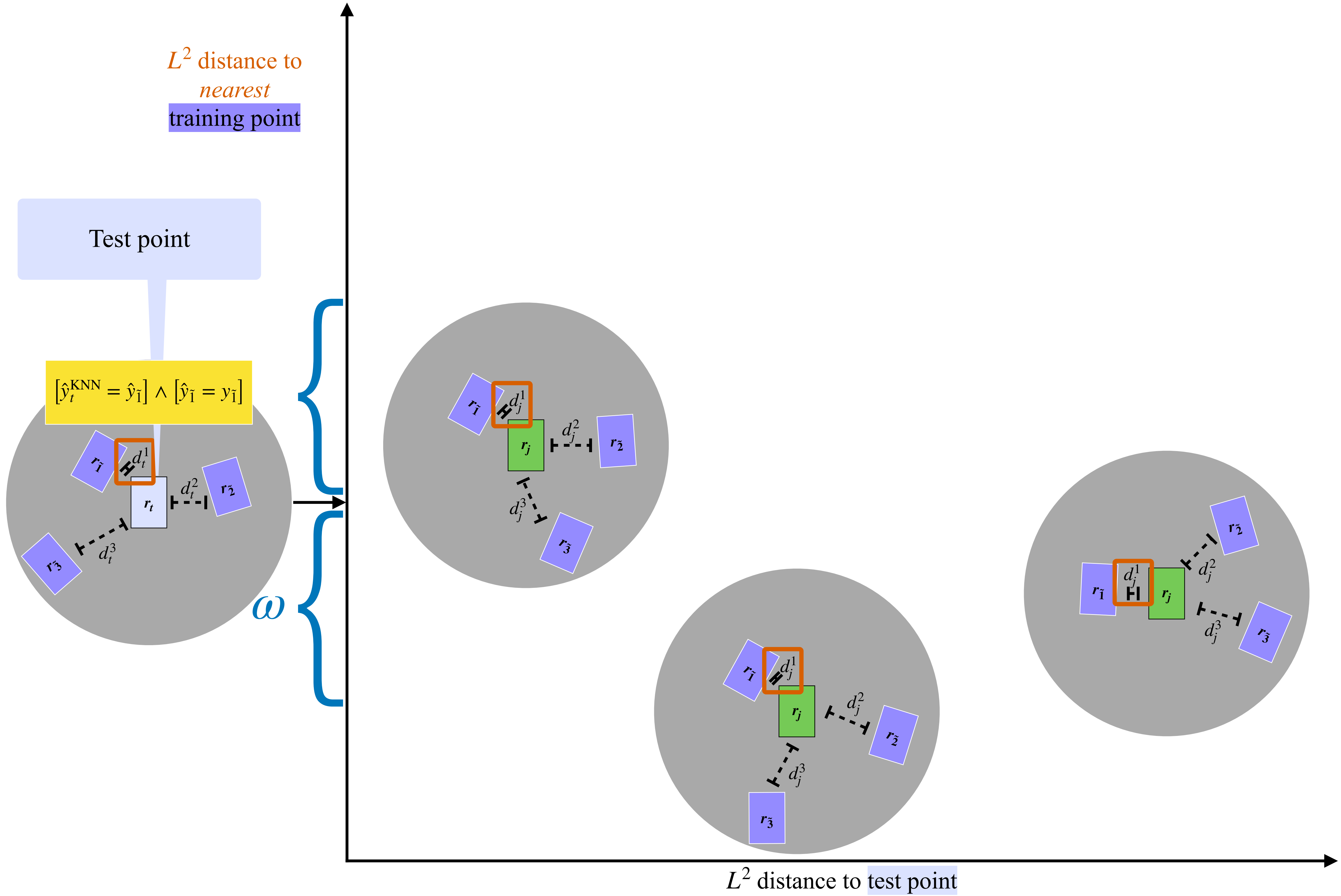
Constrain calibration points to match constraint feature



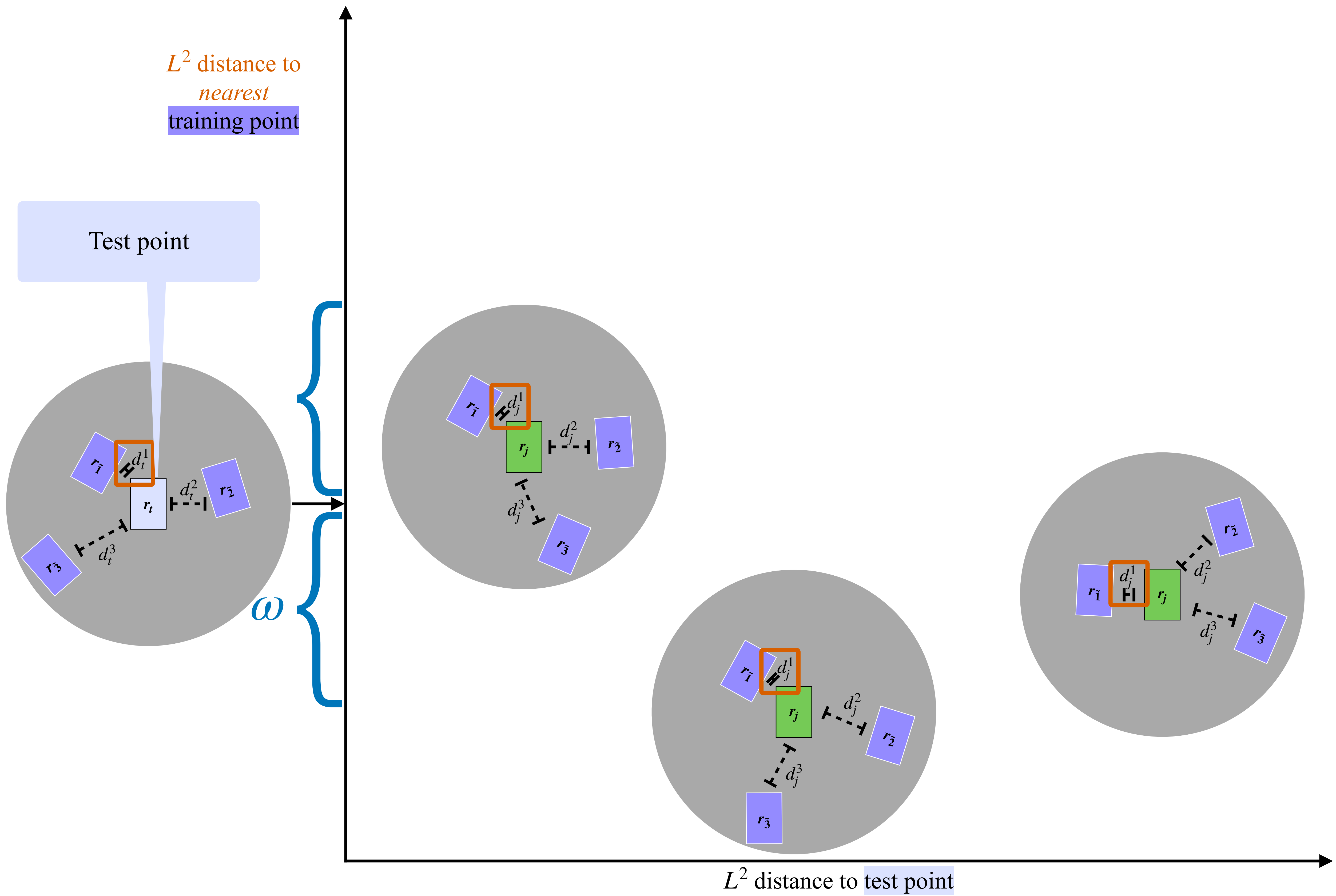
Constrain calibration points to match constraint feature



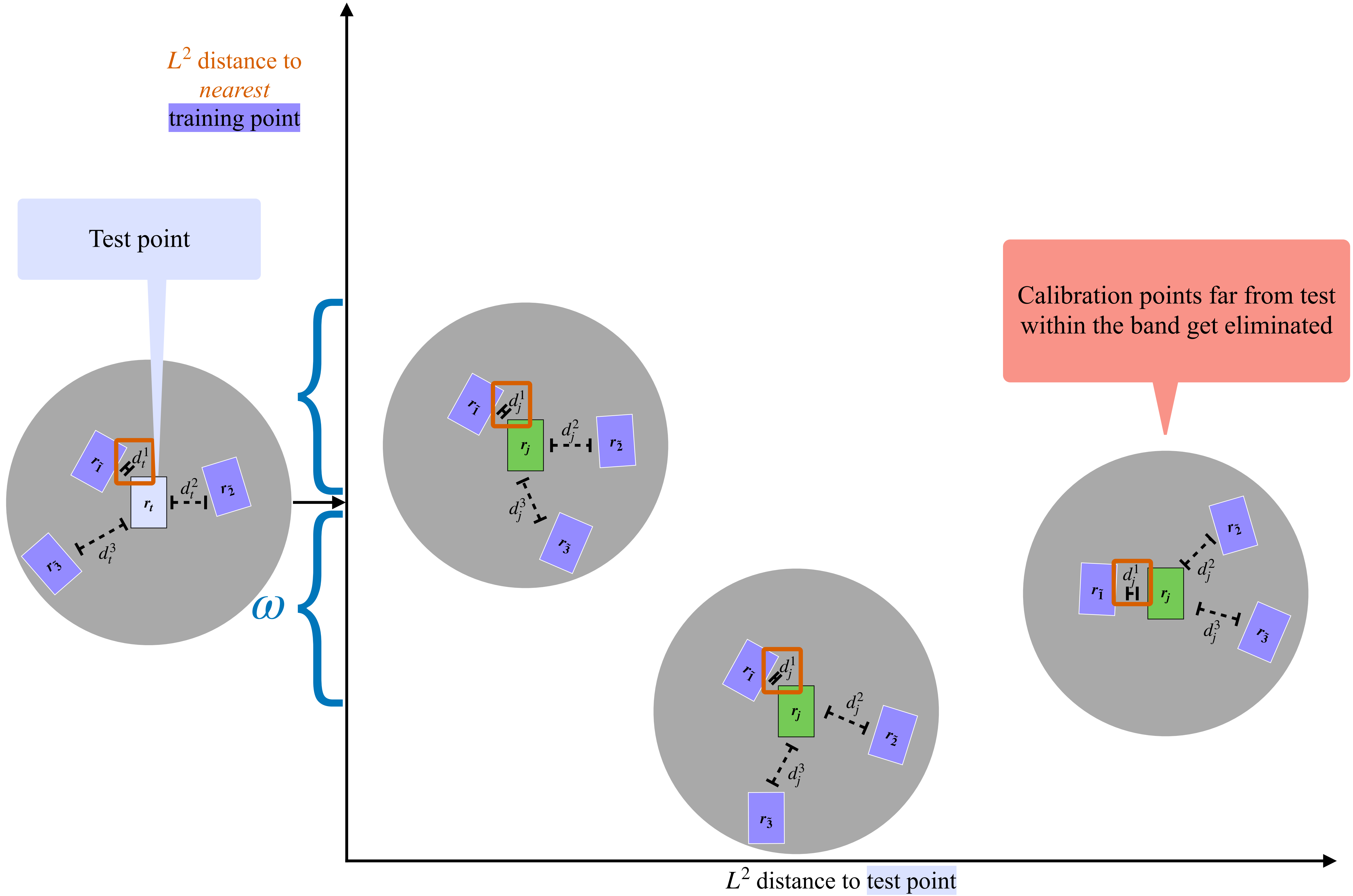
Constrain calibration points to match constraint feature



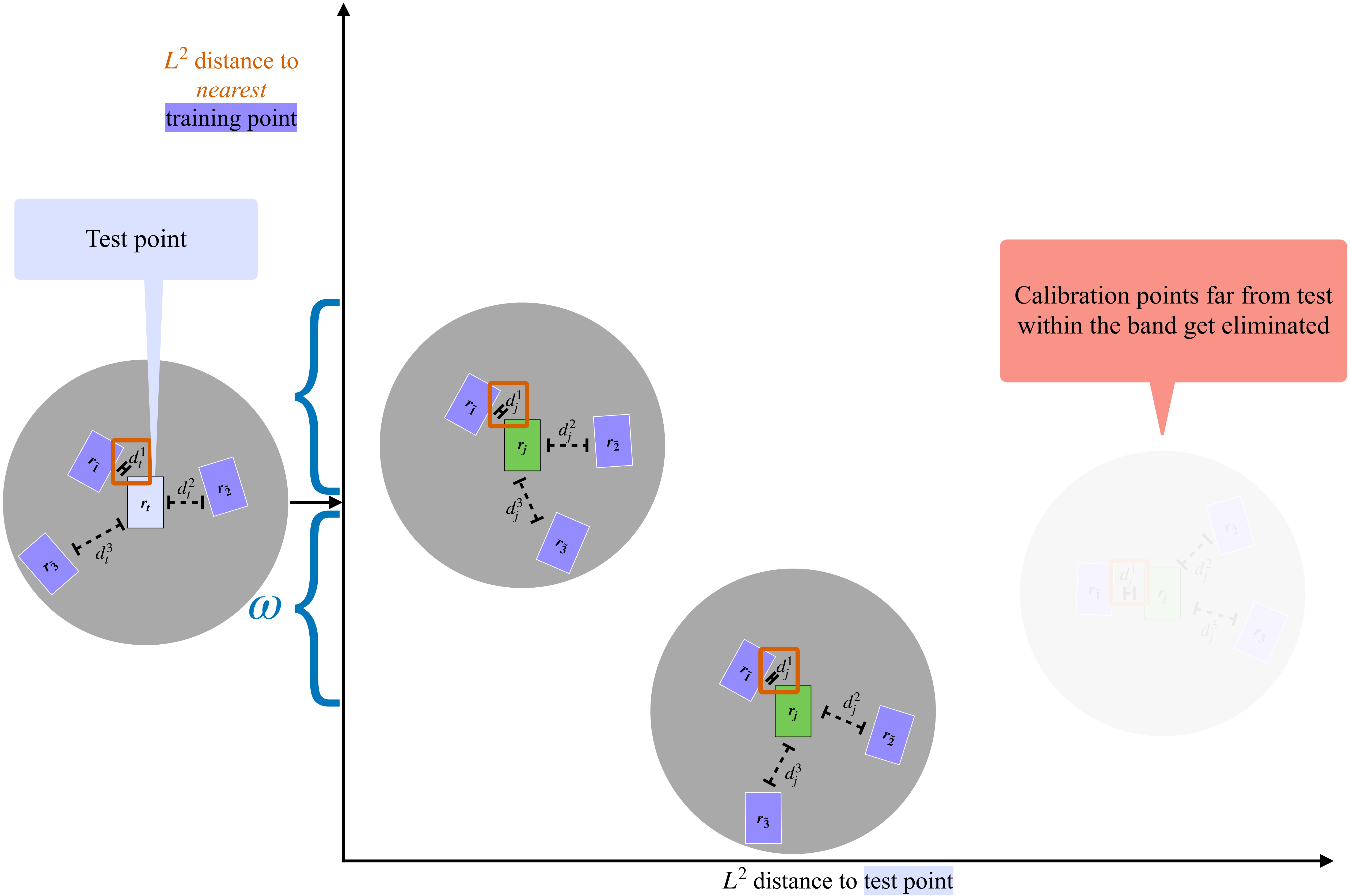
Re-sample $\mathscr{D}_{\mathrm{ca}}$ to be more similar to $\mathscr{D}_{\mathrm{te}}$



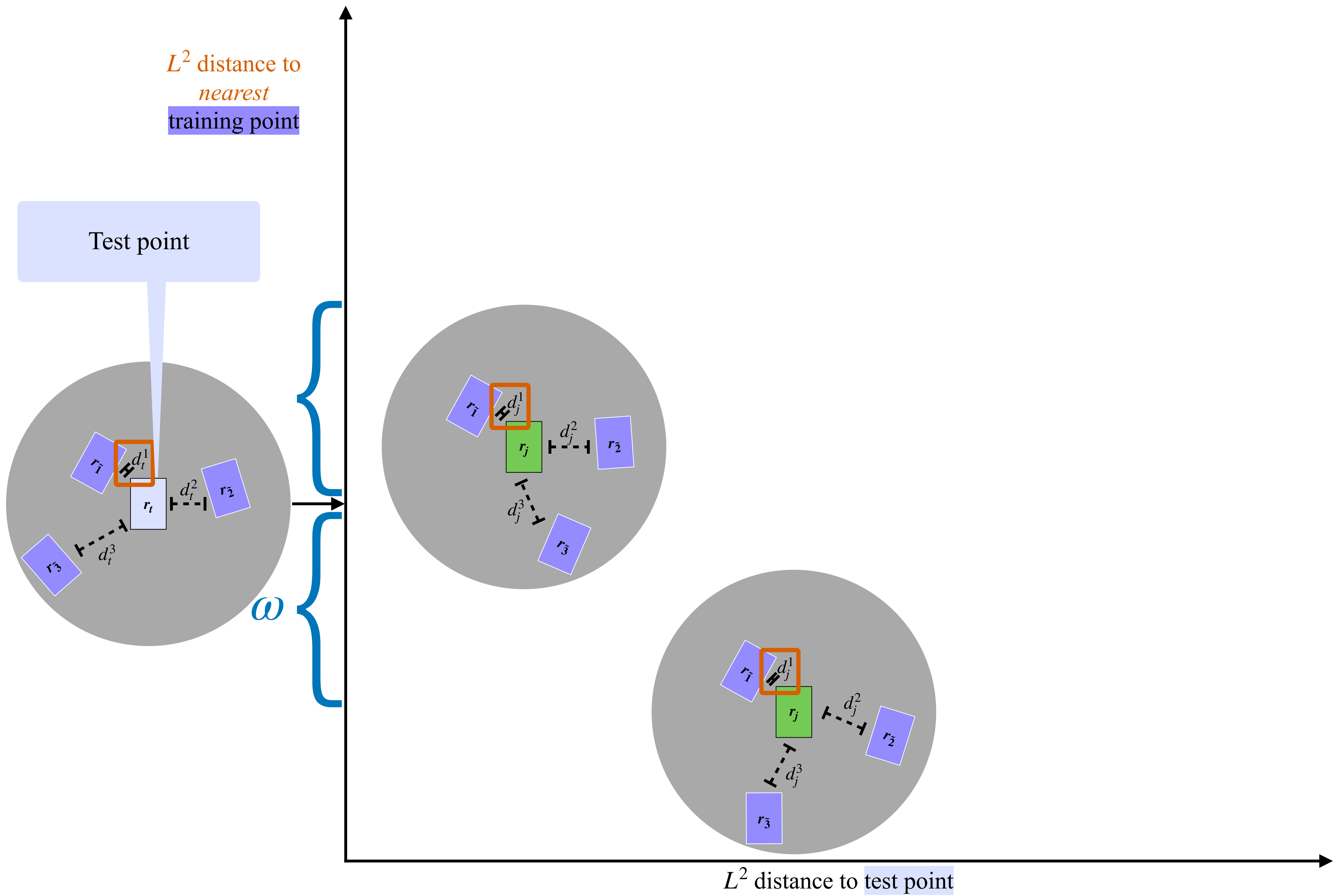
Re-sample $\mathscr{D}_{\mathrm{ca}}$ to be more similar to $\mathscr{D}_{\mathrm{te}}$



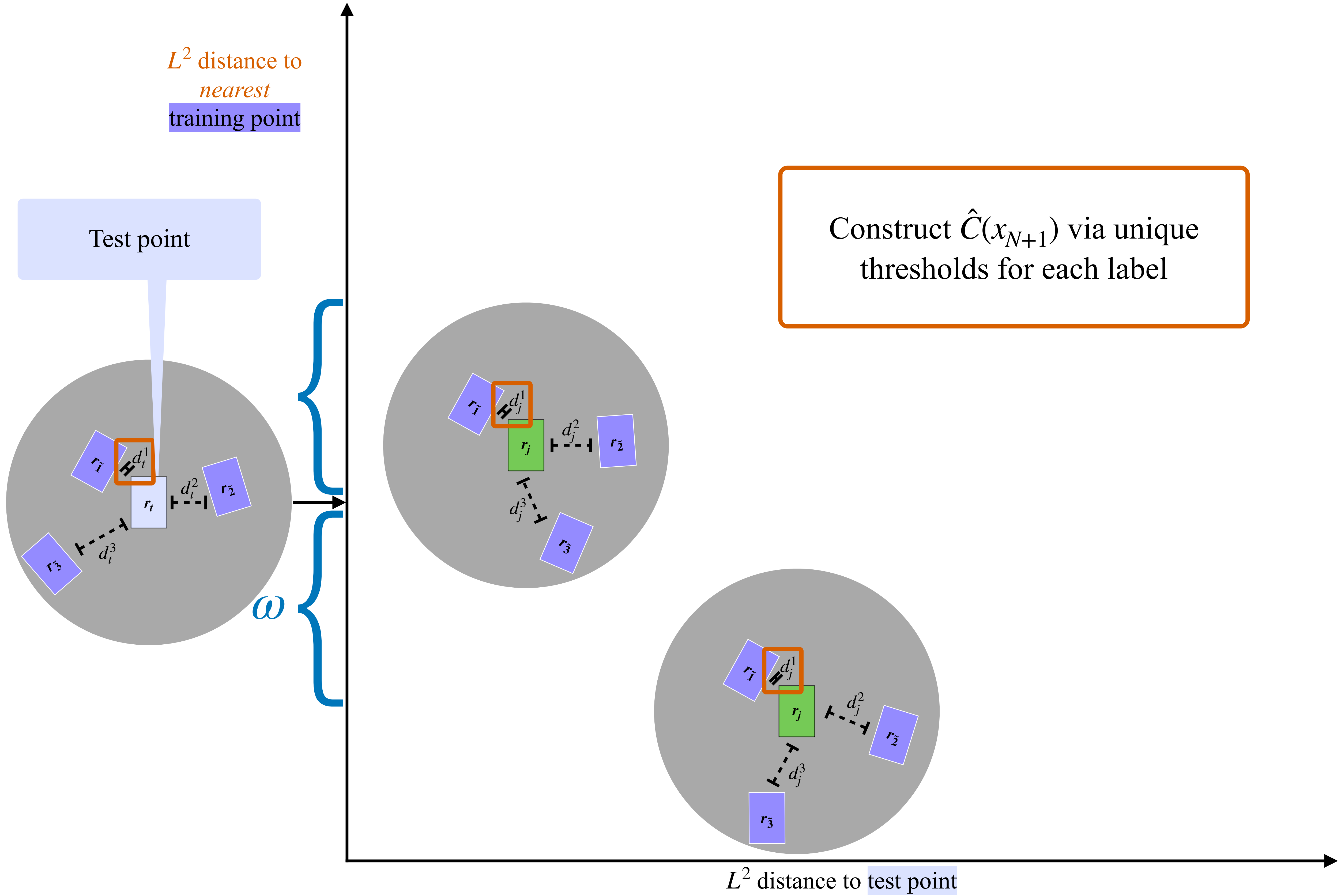
Re-sample $\mathscr{D}_{\text{ca}}$ to be more similar to $\mathscr{D}_{\text{te}}$



Re-sample $\mathscr{D}_{\text{ca}}$ to be more similar to $\mathscr{D}_{\text{te}}$

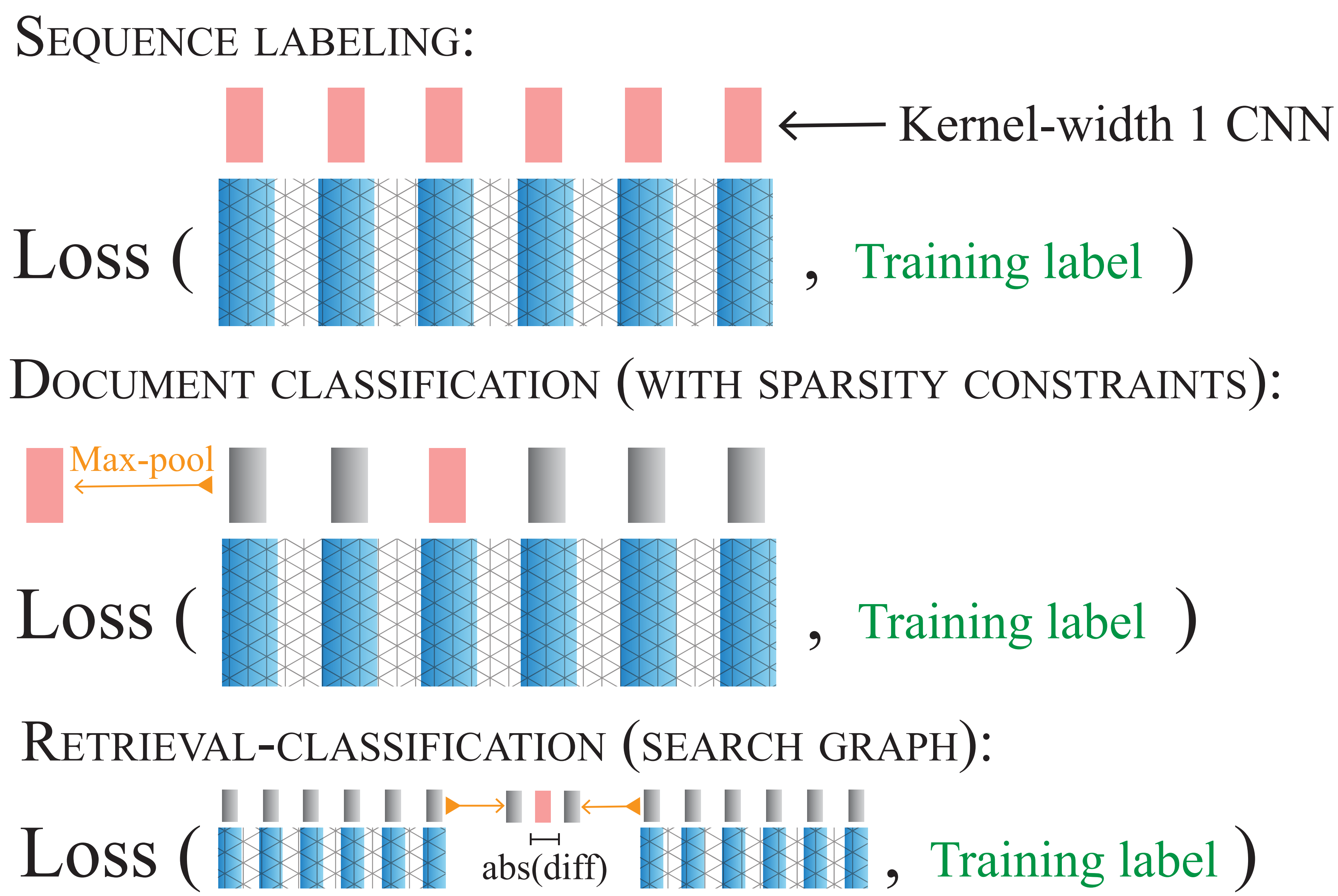


Label-conditional conformal thresholds



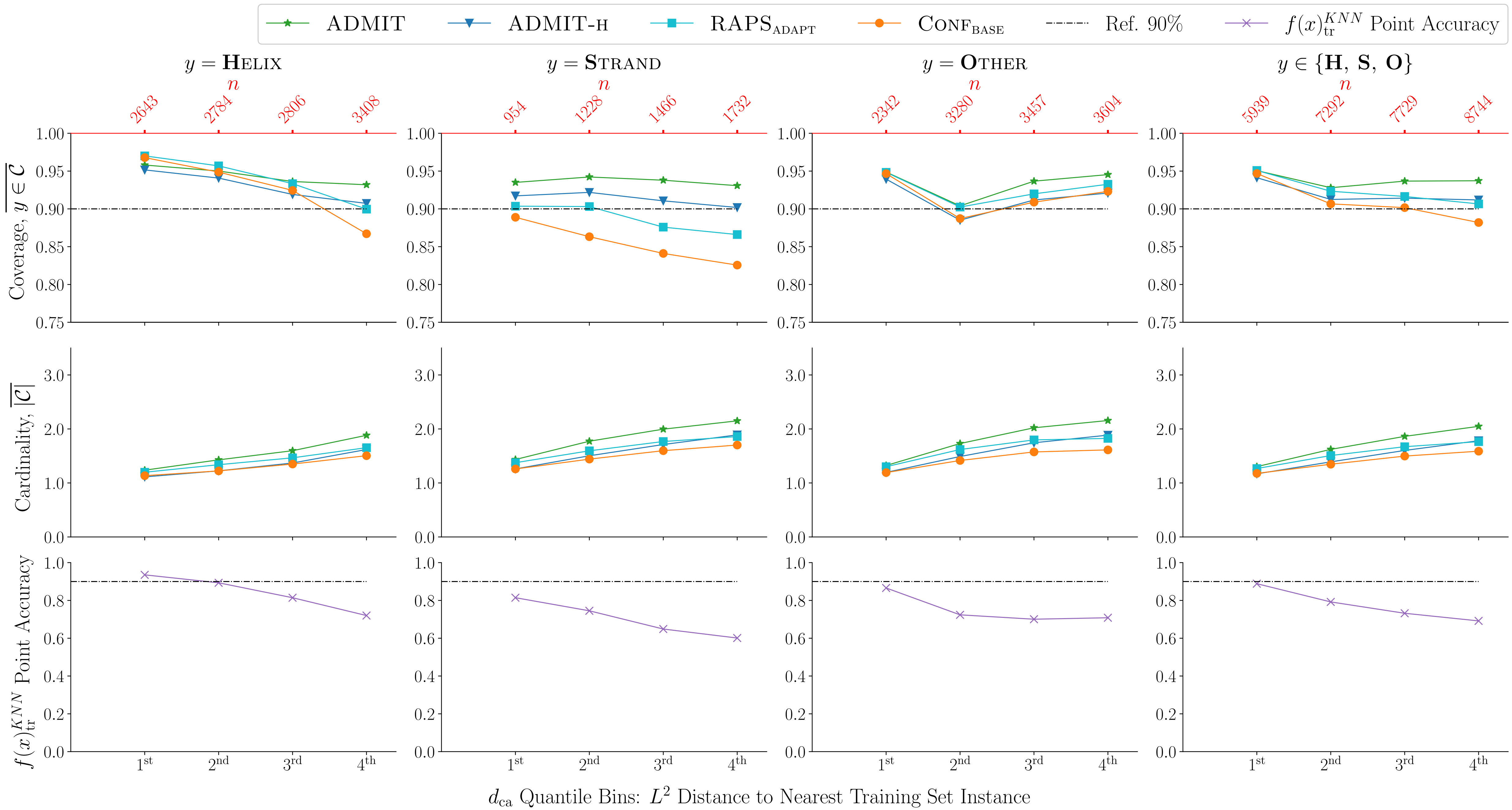
Construct $\hat{C}(x_{N+1})$ via unique thresholds for each label

Experiments



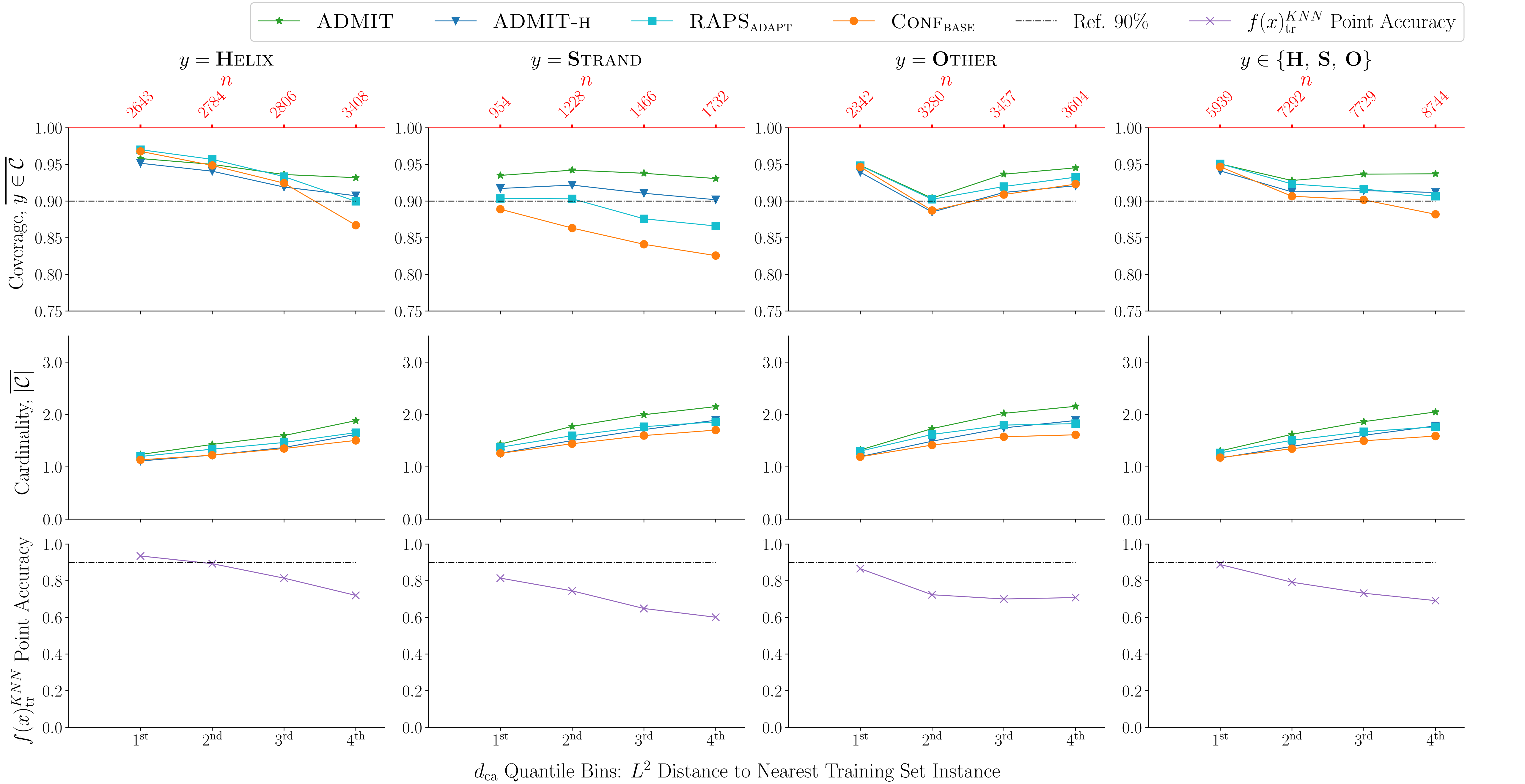
Task diversity: From in-distribution, high-accuracy *to* distribution-shifted, low-accuracy

Empirical behavior on in-domain data



Coverage, cardinality, and point accuracy for the TS115 test set from the PROTEIN task.

Empirical behavior on in-domain data



Coverage, cardinality, and point accuracy for the TS115 test set from the PROTEIN task.

Additional results

- Covariate/label shifts
- Heuristics for coverage conditioned on set composition
- And more …
- See: https://arxiv.org/abs/2205.14310

ADMIT: A general framework for constructing, constraining, and analyzing point predictions and distribution-free prediction sets for deep neural networks.